ON THE COMBINATORICS OF SAMPLE COMPRESSION SCHEMES

A Thesis

Submitted to the Faculty of Graduate Studies and Research

In Partial Fulfillment of the Requirements

for the Degree of

Masters of Science

In

Mathematics

University of Regina

By

Ryan Tessier

Regina, Saskatchewan

August 2013

© Copyright 2013: Ryan Tessier
Ryan Brett Tessier, candidate for the degree of Master of Science in Mathematics, has presented a thesis titled, *On The Combinatorics of Sample Compression Schemes*, in an oral examination held on August 23, 2013. The following committee members have found the thesis acceptable in form and content, and that the candidate demonstrated satisfactory knowledge of the subject material.

External Examiner: *Dr. Richard Anstee, University of British Columbia*

Co-Supervisor: Dr. Shaun Fallat, Department of Mathematics & Statistics

Co-Supervisor: Dr. Sandra Zilles, Department of Computer Science

Committee Member: Dr. Karen Meagher, Department of Mathematics & Statistics

Chair of Defense: Dr. Nader Mobed, Department of Physics

*Not present at defense*
Abstract

A sample compression scheme of size $k$ for a concept class $C$ is a pair of functions $(f, g)$ called the compression function and the reconstruction function. The functions have the property that for any sample $S$ consistent with a concept in $C$, $f$ compresses $S$ to some subset of $S$, for which $g$ returns a set of domain points, labelled consistently with the original sample $S$. The sample compression scheme is called labelled if the compression sets are labelled subsets of $S$ and unlabelled if the compression sets are subsets of the instance set of $S$. M. Warmuth and S. Floyd have shown that if a sample compression scheme of size equal to the VC dimension of a concept class $C$ exists then $C$ can be PAC learned by some learning algorithm. Although it is already known that any concept class of finite VC dimension is PAC learnable, the existence of a sample compression scheme of size equal to the VC dimension improves the sample complexity of learning some concept classes. This leads to an important conjecture, first proposed by M. Warmuth and S. Floyd: does there always exist a sample compression scheme of size $O(d)$ for a concept class $C$ with VC dimension $d$. 
This thesis examines a modification of sample compression schemes, specifically, for a concept class $\mathcal{C}$ we define a sequence-based sample compression scheme for $\mathcal{C}$ as a pair of functions $(f^*, g^*)$ where the items we compress to are now sequences instead of sets. Here we can differentiate between labelled and unlabelled sequence-based sample compression schemes in a similar fashion as with standard sample compression schemes. We look at properties of sequence-based sample compression schemes and also discuss a few sequence-based sample compression scheme algorithms and determine how they improve compression bounds over the original set-based compression scheme algorithms. Finally we discuss connections between set and sequence-based sample compression schemes and design theory.
Acknowledgements

I would like thank the University of Regina Faculty of Graduate Studies for financial support via the Centennial Merit Scholarship. Further I would like to thank my supervisors Dr. S. Fallat and Dr. S. Zilles for all that they have taught me and for their time and support. Finally I would like to express my thanks to my thesis examiners Dr K. Meagher and Dr. R. Anstee for their reading of my thesis, as well as the improvements they suggested.
Contents

Abstract i

Acknowledgements iii

Table of Contents iv

List of Figures vi

List of Tables vi

1 Introduction 1

1.1 Organization .......................................................... 4

2 Background and Motivation 6

2.1 Basic Definitions ..................................................... 6

2.2 VC Theory ............................................................ 9

2.3 Labelled Sample Compression Scheme ............................ 15

2.4 Unlabelled Compression Schemes ................................. 19
List of Figures

4.1 The Fano Plane ........................................ 61

List of Tables

2.1 The incidence matrix for the concept class $C$ ...................... 8
2.2 $I_C$ restricted to $x_4$ and $x_5$ ................................ 11
2.3 Maximum concept class of VC dimension 2 ......................... 33
2.4 Representation mappings for a maximum class of VC dimension 2 . 34
2.5 Standard Maximum Class of VC dimension 2 ....................... 35

4.1 The incidence matrix of the Fano plane ............................. 63
4.2 The restriction to $x_1$ and $x_2$ of the incidence matrix for the Fano plane 64
Chapter 1

Introduction

In computational learning theory the goal is to create ‘smart’ learners. By a ‘smart’ learner, we mean one that learns a given concept with as few examples as possible. Formalized, we consider a concept to be a set $C \in \mathcal{C}$ where $\mathcal{C} \subseteq 2^X$ is called a concept class and $X$ is some set which is referred to as an instance space. A learner then receives some set of points from $X$ labelled with either a 1 or a 0, called examples, based on the inclusion of points from $X$ in $C$. The learner’s job is to then return some binary function, called a hypothesis, that has points from $X$ labelled according to $C$. The details such as what examples are received by the learner, how they are given to the learner and the measurement of how well the hypothesis predicts the concept $C$ from the examples given to it are all things that depend on the model of learning that is used.

The probably approximately correct (PAC) learning model is a learning model
proposed by Leslie Valiant in [3]. Here a concept class $C \subseteq 2^X$ is called **PAC learnable** if we can find some function $L$ (called a **learning algorithm**, or **learning rule**) such that for any $0 < \epsilon, \delta \leq 1$ and for any probability distribution $D$ there is some number $n$ (called the **sample complexity** of $L$) which depends on $\epsilon$ and $\delta$, such that if more than $n$ examples drawn at random from $D$ and provided to our learning algorithm, the probability that the algorithm makes more than $\epsilon$ errors is less than the quantity $\delta$.

With PAC learning defined this way, Blumer et al. showed in [7] that the PAC learnability of a concept class $C$ depends entirely on whether the Vapnik-Chervonenkis (VC) dimension of $C$ is finite. This is a fundamental result in PAC learning theory as it relates PAC learnability to a purely combinatorial property of concept classes. Blumer et al. also demonstrated that the sample complexity of a learning algorithm that PAC learns a concept class is linear in the VC dimension, where the sample complexity of a learning algorithm is just the number of samples required by the algorithm to learn a given concept.

Sample compression schemes were first introduced by Littlestone and Warmuth in [4] as an alternative way to characterize PAC learnability. A sample compression scheme for some concept class $C$ is a pair of functions $(f, g)$. The function $f$ is called the compression function. This function takes, as input, a set of labelled examples and returns a subset of these examples. The function $g$ is called the reconstruction
function. This function takes the subset produced by $f$ and returns a hypothesis that is consistent with some concept class. Here a labelled example is just a set of domain points with some binary labelling and a hypothesis is just binary function.

In [4] it was shown that the existence of a sample compression scheme of a particular size for a concept class $C$ was sufficient to ensure that the class was PAC learnable. This work was extended in [11] by Floyd and Warmuth showing that not only does a sample compression scheme ensure PAC learnability, but that the sample complexity is linear in the size of the sample compression scheme. Further they observed that for a number of very natural examples of concept classes, they could find sample compression schemes of size equal to the VC dimension. Then in [18] Kuzmin and Warmuth showed that this property holds for the same type of concept classes even when we restrict our compression sets to being unlabelled subsets of our labelled examples. This suggests that there may be a deep mathematical connection between sample compression schemes (both labelled and unlabelled), PAC learnability and the VC dimension.

This thesis looks at another modification of the original definition of sample compression schemes. Here we consider the case where our compression objects are sequences of points from the input sample set instead of sets. The immediate advantage of this approach is that when we consider sequences, we expand the number of objects that we can compress samples to since we may consider sequences as ‘copies’
of compression sets. We also have an advantage relative to the sample set we are drawing points from. For any sample set $S$ there are $\binom{|S|}{k} k!$ unlabelled or labelled subsequences of size $k$; however for labelled and unlabelled compression sets there are only $\binom{|S|}{k}$ subsets of $S$ that have size $k$. The work of this thesis primarily makes use of this fact.

1.1 Organization

Chapter 2 describes the basic definitions and results needed to understand the main work of this thesis. This includes the notion of concept classes, the VC dimension, basic PAC learning theory, and how sample compression schemes relate to PAC learning. Chapter 3 gives a general definition of sample compression schemes and we will use this definition to define the notions of labelled and unlabelled sequence-based sample compression schemes. Additionally it compares this definition to another generalization of sample compression schemes. Chapter 4 describes the majority of the mathematical results on sequence based sample compression. This chapter will present a few new unlabelled sequence-based sample compression schemes: these include an unlabelled sequence-based sample compression scheme of size $\lceil \frac{m^2}{2} \rceil$, where $m$ is the size of the instance space, and unlabelled sample compression scheme that has size less than the VC dimension for finite cases. We also establish bounds on the minimum sizes of labelled sequence-based sample compression schemes. Later in
this chapter we discuss the compression of combinatorial $t$-designs, finding that for $\lambda$
small enough we can find an unlabelled sequence-based sample compression scheme
of size $t$, and further we observe that it allows us to compute the VC dimension of
Steiner systems.
Chapter 2

Background and Motivation

We begin by giving some background on sample compression schemes as well as PAC learnability as a way to give context to sequence-based compression along with motivating our approach. As a matter of convention, from here on we will assume that \( N = \{0, 1, 2, 3, \ldots \} \) and that \( Z^+ = N \setminus \{0\} \). We will also let \([n] = \{1, 2, \ldots, n\} \subseteq Z^+\) for any \( n \in Z^+\).

2.1 Basic Definitions

**Definition 2.1.1.** Let \( X \) be an arbitrary set which will be referred to as an instance space and let \( C \subseteq X \). Here we call \( C \) a concept and we call any collection of concepts a concept class and denote this by \( \mathcal{C} \), note that \( \mathcal{C} \subseteq 2^X \).

Any \((x, y) \in X \times \{0, 1\}\) is called an example or a labelled example. Further, any non-empty subset \( S \subseteq X \times \{0, 1\}\) is called a sample set.
Definition 2.1.2. For a concept \( C \in 2^X \) define a function \( C : X \to \{0, 1\} \) such that for all \( x \in X \)

\[
C(x) = \begin{cases} 
1, & \text{if } x \in C; \\
0, & \text{if } x \notin C.
\end{cases}
\]

We call this function the **characteristic function** of the concept \( C \).

This characteristic function allows us to think of any \( C \subseteq 2^X \) as both a collection of subsets over \( 2^X \) and as a collection of binary functions over the domain \( X \). This view of characteristic functions will allow us to define the notion of an incidence matrix of a concept class as well as to relate sample sets to concept classes.

Definition 2.1.3. For \( n, m \in \mathbb{N} \) and a concept class \( \mathcal{C} = \{C_1, C_2, \ldots, C_n\} \subseteq 2^X \) where \( X = \{x_1, x_2, \ldots, x_m\} \), define the **incidence matrix** of \( \mathcal{C} \) to be the binary \( n \times m \) matrix \( I_{\mathcal{C}} \) given by

\[
I_{\mathcal{C}} = [C_1(X)C_2(X)\ldots C_n(X)]^T.
\]

The \( i \)-th column of \( I_{\mathcal{C}} \) is labelled \( x_i \in X \), and the \( j \)-th row of \( I_{\mathcal{C}} \) is labelled \( C_j \in \mathcal{C} \), and the \((i, j)\)-entry is 1 if \( x_i \in C_j \) and 0 otherwise.

Example 2.1.4. Assume that \( \mathcal{C} = \{\{x_1, x_3, x_5\}, \{x_2, x_4\}, \{x_2\}, \{x_4, x_5\}\} \subseteq 2^X \) where \( X = \{x_1, x_2, x_3, x_4, x_5\} \). Then the incidence matrix of \( \mathcal{C} \) is \( I_{\mathcal{C}} \) where \( I_{\mathcal{C}} \) is given by Table 2.1.
Table 2.1: The incidence matrix for the concept class $\mathcal{C}$

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$C_2$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$C_3$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Definition 2.1.5.** Let $S \subseteq X \times \{0, 1\}$ be a sample set. We say that $S$ is labelled consistently with a concept $C \in \mathcal{C}$ if for any $(x, y) \in S$ we have that $y = C(x)$. The notation $S_C$ refers to a sample that is labelled consistently with a concept $C \in \mathcal{C}$.

While, using this notation, $S_C$ may not be a unique set there are conditions under which $S_C$ is a unique set, namely if $|S_C| = |X|$. This definition also means that if $S \subseteq 2^{X \times \{0, 1\}}$ is a sample labelled consistently with some concept $C \in \mathcal{C}$, there is no $x \in X$ such that both $(x, 1), (x, 0) \in S$. These facts will help us in formally defining compression schemes later in the thesis.

Given a sample set $S \subseteq 2^{X \times \{0, 1\}}$ it is often useful to refer to the set of points from $X$ that are given labels in $S$.

**Definition 2.1.6.** Let $S \subseteq X \times \{0, 1\}$ be a set of labelled examples. We define the
instance set of $S$ to be the set

$$X(S) = \{ x \in X : (x,y) \in S \}.$$  

We can also define a couple of sets that are related to the notion of an instance set, but describe only the domain points with a given labelling. Before we define the primary algorithms of this section, some shorthand will be necessary to define for ease of exposition.

**Definition 2.1.7.** If $S \subseteq 2^{X \times \{0,1\}}$ then we define

$$1(S) = \{ x \in X(S) | (x,1) \in S \},$$

and

$$0(S) = \{ x \in X(S) | (x,0) \in S \}.$$  

Additionally, for a concept class $C \subseteq 2^X$ there is some subset $A \subseteq X$ such that if $C \in C$ then $C \subseteq A$, the minimal such set $A$ is helpful to define.

**Definition 2.1.8.** If $C \subseteq 2^X$ then we define the domain of $C$ to be the set

$$\text{Dom}(C) = \{ x \in X : x \in C \text{ for some } C \in C \}.$$  

2.2 VC Theory

This section focuses on defining the combinatorial parameter known as the Vapnik-Chervonenkis (VC) dimension along with some basic results on this quantity. We
also define some special concept classes that have a structure dependent on the VC dimension.

**Definition 2.2.1.** Given a concept class $\mathcal{C}$ over an instance space $X$, and a subset $A \subseteq X$, we define the restriction of $\mathcal{C}$ to $A$ to be the set $\mathcal{C}|_A := \{ C \cap A : C \in \mathcal{C} \}$.

Here we do not allow the restriction $\mathcal{C}|_A$ to be a multi-set. That is, if there are two elements $C_1, C_2 \in \mathcal{C}$ such that $C_1 \cap A = C_2 \cap A$ then $C_1 \cap A \in \mathcal{C}|_A$ while $C_2 \cap A \notin \mathcal{C}|_A$.

The notion of the restriction of a concept class is used in the definition of shattering.

**Definition 2.2.2.** If $\mathcal{C}$ is a concept class over an instance space $X$, we say that a subset $A \subseteq X$ is **shattered** by $\mathcal{C}$ if $\mathcal{C}|_A = 2^A$.

**Definition 2.2.3.** Define the **Vapnik-Chervonenkis dimension**, or VC dimension, of a class $\mathcal{C} \subseteq 2^X$ to be the quantity

$$\text{VCD}(\mathcal{C}) := \sup \{|A| : A \subseteq X, A \text{ is finite, and } A \text{ is shattered by } \mathcal{C}\}.$$  

Based on this definition we can see the VC dimension of a concept class $\mathcal{C}$ is just the size of the largest subset of $X$ that is shattered by $\mathcal{C}$.

**Example 2.2.4.** To illustrate the idea of shattering and the VC dimension let $X = \{x_1, x_2, x_3, x_4, x_5\}$ and let $\mathcal{C}$ be the concept class $\mathcal{C}$ that corresponds to the incidence matrix $I_\mathcal{C}$ in Table 2.1. Here we see that $\mathcal{C}|_{\{x_4,x_5\}} = 2^X$ and $I_\mathcal{C}|_{\{x_4,x_5\}}$ is given by Table 2.2.
\begin{array}{cc}
\begin{bmatrix}
0 & 1 \\
1 & 0 \\
0 & 0 \\
1 & 1 \\
\end{bmatrix}
\end{array}

Table 2.2: \(I_C\) restricted to \(x_4\) and \(x_5\)

So \(C\) shatters a set of size 2. However, we notice that since \(|C| < 2^3 = 8\) we cannot possibly shatter sets of any size larger. Therefore we say that \(C\) has VC dimension 2 and write \(VCD(C) = 2\).

For \(m, d \in \mathbb{N}\) with \(d \leq m\) let \(\binom{m}{\leq d}\) denote

\[
\binom{m}{\leq d} = \sum_{i=0}^{d} \binom{m}{i}.
\]

By convention, if \(d > m\) we will define

\[
\binom{m}{\leq d} := 2^m.
\]

Similarly for a set \(A\) and \(d \in \mathbb{N}\) we let

\[
\binom{A}{d} := \{B \subseteq A : |B| = d\}.
\]

Furthermore, we let

\[
\binom{A}{\leq d} := \{B \subseteq A : |B| \leq d\}.
\]
It is easy to check for an instance space $X$ with $|X| = m \in \mathbb{N}$ and for some $d \in \mathbb{N}$ with $d \leq m$ that $\binom{|X|}{d} = \binom{m}{d}$.

Using this notation, we can obtain a combinatorial inequality that will be useful in bounding sequence-based sample compression schemes later in this work.

**Theorem 2.2.5.** [1] For all $m, d \in \mathbb{N}$

$$\binom{m}{d} \leq \left( \frac{em}{d} \right)^d.$$

The next theorem is a fundamental result of VC theory. It predicts an upper bound on the size of a concept class of a particular VC dimension. In addition to allowing for the definition of maximum concept classes; it also allows for a way to distinguish between maximum and maximal classes of some VC dimension.

**Theorem 2.2.6** (Sauer’s Lemma). [1] Let $C \subseteq 2^X$ be a finite concept class of VC dimension $d$ over a finite instance space $Y$ with $|Y| = m$ and $Y \subseteq X$. Then

$$|C| \leq \sum_{i=0}^{d} \binom{m}{i}.$$

**Definition 2.2.7.** Let $X$ be some instance space. We call $C \subseteq 2^X$ a maximum class of VC dimension $d$ if for any finite subset $Y \subseteq X$,

$$|C|_Y = \binom{|Y|}{\leq d}.$$

Although the previous definition may seem somewhat convoluted, it is necessary to be able to define maximum classes over infinite instance spaces. The following
Theorem 2.2.8. [5] Let $\mathcal{C} \subseteq 2^X$ be a concept class of VC dimension $d$ on a finite instance space $X$ with $|X| = m$. The class $\mathcal{C}$ is maximum if and only if

$$|\mathcal{C}| = \sum_{i=0}^{d} \binom{m}{i}.$$ 

Next we define a very natural example of a maximum class. In some sense this is the most natural maximum class as it is exactly the set of all sets up to a particular size.

Definition 2.2.9. For $d \in \mathbb{N}$ define $\mathcal{K}_d = \{ C \in 2^X : |C| \leq d \}$. We call this concept class the **standard maximum class of VC dimension** $d$.

Proposition 2.2.10. For any $d \in \mathbb{N}$ the concept class $\mathcal{K}_d$ is a maximum class with $\text{VCD}(\mathcal{K}_d) = d$.

Proof. For $d \in \mathbb{N}$ by definition $\mathcal{K}_d \subseteq 2^X$ shatters any set of $d$ points from $X$. Thus $\text{VCD}(\mathcal{K}_d) \geq d$. To show that $\text{VCD}(\mathcal{K}_d) = d$ we only need to show that $\mathcal{K}_d$ cannot shatter any set of $d + 1$ points from $X$. This is straightforward to show since to shatter $d + 1$ points $D$ from $X$ one would need to have $D \in \mathcal{K}_d$. However no set in $\mathcal{K}_d$ has more than $d$ points. Therefore $\text{VCD}(\mathcal{K}_d) = d$. \qed
Example 2.2.11. [9] Although the motivating example in the definition of a maximum class of VC dimension $d$ is that of the standard maximum class, we certainly can define other maximum classes that are not the standard maximum class. Let $X$ be an instance space with $|X| = 5$ and let $\mathcal{C} \subseteq 2^X$ be the maximum concept class of VC dimension $2$ defined by the incidence matrix $I_{\mathcal{C}}$ given by the matrix in Table 2.3 on page 33.

By Theorem 2.2.8 and since $\binom{5}{2} + \binom{5}{1} + \binom{5}{0} = 16$ we can deduce that this is a maximum class of VC dimension 2.

If $\mathcal{C}$ is a maximum class of VC dimension $d$ then adding another concept to $\mathcal{C}$ necessarily increases the VC dimension of $\mathcal{C}$. While this property is one that holds for all maximum classes, it is not a property that is unique to maximum classes. There exist classes $\mathcal{C}'$ with $|\mathcal{C}'| < \sum_{i=0}^{d} \binom{m}{i}$ such that if any concept is added to $\mathcal{C}'$, then the VC dimension of $\mathcal{C}'$ increases.

Definition 2.2.12. Let $\mathcal{C} \subseteq 2^X$ be a concept class. If the VC dimension of $\mathcal{C}$ is $d$, then $\mathcal{C}$ is called maximal of VC dimension $d$ if adding any concept to $\mathcal{C}$ increases the VC dimension of $\mathcal{C}$.

Maximum classes are defined by their size, however maximal classes can vary greatly in size. It is a natural question to ask if there is some lower bound on the size of maximal classes. The next theorem provides such a bound.
Theorem 2.2.13. [26] If $\mathcal{C} \subseteq 2^X$ is a maximal class with $\text{VCD}(\mathcal{C}) = d \in \mathbb{N}$, then

$$|\mathcal{C}| \geq 2^m - 2^{m-d-1}(\frac{m}{d+1}).$$

2.3 Labelled Sample Compression Scheme

The notion of sample compression schemes was introduced in [4] by Warmuth and Littlestone. In [11] Floyd and Warmuth expanded on the work done in [4] by providing a sample compression scheme for maximum classes that is linear in the VC dimension. They went on to demonstrate that if a concept class has a sample compression scheme of a size that is linear in the VC dimension, then the concept class must be PAC learnable by an algorithm with a sample complexity that is also linear in the VC dimension. The sample complexity obtained improved marginally on already known bounds for PAC learning for a variety of concept classes.

This section will define sample compression and survey some results on the constructability of sample compression schemes as well as discuss the connections between sample compression and the VC dimension.

Definition 2.3.1 (Sample Compression Scheme). Let $X$ be an instance space and let $\mathcal{C} \subseteq 2^X$ be a concept class. A sample compression scheme or a SCS of size at most $k$ for $\mathcal{C}$ is a pair $(f, g)$ where $f : 2^X \times \{0,1\} \rightarrow 2^X \times \{0,1\}$ and $g : 2^X \times \{0,1\} \rightarrow 2^X \times \{0,1\}$ such that if $S_C \subseteq X \times \{0,1\}$ is a sample labelled consistently with some $C \in \mathcal{C}$, then

i) $f(S_C) \subseteq S_C$, 

15
ii) \( |f(S_C)| \leq k \),

iii) \( S_C \subseteq g(f(S_C)) \),

iv) \( g(f(S_C)) = S_{C'} \) for some \( C' \subseteq X \) and \( |S_{C'}| = |X| \).

We say that \((f, g)\) has size \( k' \) if \( k' \) is the smallest \( k \) for which \((f, g)\) is a sample compression scheme of size at most \( k \). Here we call \( f(S_C) \) a compression set.

The classic example of a sample compression scheme comes from the plane \( \mathbb{R}^2 \). In this work we only consider the finite case because Theorem 2.5.8 will guarantee that to prove compression scheme existence, we only need to worry about the case when the underlying instance space is finite. However, motivation for defining sample compression schemes comes in part from the observation that sample compression relates nicely to the VC dimension of a concept class when the objects we are trying to compress are axis-parallel rectangles in the plane.

**Example 2.3.2.** We let \( \mathcal{R} \) be the class of axis-parallel rectangles in \( X = \mathbb{R}^2 \). Every concept \( R \in \mathcal{R} \) corresponds to an axis-parallel rectangle in \( \mathbb{R}^2 \) with points \( x \in X \) labelled 1 if \( x \in R \) and labelled 0 if \( x \notin R \).

In this setting if \( S \subseteq \mathbb{R}^2 \times \{0, 1\} \) is a sample set consistent with some \( R \in \mathcal{R} \), a valid sample compression scheme of size at most 4 for \( \mathcal{R} \) consists of a compression function that, for a sample \( S \) with \( |\mathbf{1}(S)| \geq 4 \), takes the topmost, bottommost, leftmost and rightmost positively labelled examples in \( S \) and compresses to this set.
of points each with label 1, call this set $D$. If $|1(S)| < 4$ then the compression function just compresses to the set $\{(x,1)|x \in 1(S)\}$. The reconstruction function $g$ returns the smallest axis aligned rectangle $R' \in \mathcal{R}$ that is consistent with the compression set $D$. So we see that the properties (i)-(iv) in Definition 2.3.1 are satisfied as a sample compression scheme since

i) $f(S) = D \subseteq S$

ii) $|f(S)| \leq 4$

iii) $S \subseteq g(f(S)) = g(D) = R'$

iv) This property holds trivially since no point inside an axis-aligned rectangle can be labelled anything other than 1 and no point outside an axis-aligned rectangle can be labelled anything other than 0.

The size of this compression scheme is interesting since there exists some set of 4 points in the plane for which we can produce all possible labellings of 0’s and 1’s by choosing different axis-aligned rectangles to fit around those points, so it can be said that $\mathcal{R}$ shatters a set of 4 points. However there is no set of 5 points in the plane that $\mathcal{R}$ can shatter, therefore $VCD(\mathcal{R}) = 4$. It is interesting to note that the VC dimension of this class is exactly the same as the size of the previous compression scheme for this class. This suggests that there may be some connection between the size of a sample compression scheme for a concept class $C$ and the VC dimension of
A basic result on sample compression schemes is that every finite concept class has a sample compression scheme of some size. Furthermore, the upper bound on the size of the compression scheme is logarithmic in the size of the concept class.

**Theorem 2.3.3.** [11] Let $C \subseteq 2^X$ be a finite concept class. Then there exists a labelled sample compression scheme of size at most $\log |C|$ for the class $C$.

One motivating factor in the study of sample compression schemes is the connection between the size of a sample compression scheme for a concept class and the VC dimension of the class. In [11] it was first observed that for a maximum concept class a sample compression scheme of size equal to the VC dimension could be constructed.

**Theorem 2.3.4.** [11] Let $C \subseteq 2^X$ be a maximum class of VC dimension $d$ on an arbitrary domain $X$. Then there exists a sample compression scheme of size $d$ for $C$.

As it turns out, the VC dimension also provides an asymptotic lower bound on the size of sample compression schemes for maximum classes.

**Theorem 2.3.5.** [11] If $C \subseteq 2^X$ is a maximum concept class of VC dimension $d > 0$, then there is no sample compression scheme of size less than $d$ for sample sets of size $m$ where $m \geq d^22^{d-1}$.

While the size of sample compression schemes for maximum classes is bounded
below by the VC dimension, we can also find a lower bound for the size of sample compression schemes that is similarly linear in the VC dimension.

**Theorem 2.3.6.** [11] Let $\mathcal{C} \subseteq 2^X$ be a concept class of VC dimension $d$. Then there is no sample compression scheme of size at most $\frac{4d}{5}$ for sample sets of size at least $d$.

### 2.4 Unlabelled Compression Schemes

Unlabelled sample compression schemes were introduced by Kuzmin and Warmuth in [18] as a modification of the original notion of sample compression schemes. Previous modifications of sample compression schemes often expanded the number of compression sets available. Here the objects that serve as compression sets are now unlabelled subsets of the instance sets of samples consistent with the concept class we wish to compress. It follows that in total there are now fewer compression sets than with the labelled sample compression schemes. Even so, we still find that there is a close connection between this type of compression and the VC dimension.

**Definition 2.4.1** (Unlabelled Sample Compression Scheme). Let $X$ be an instance space and let $\mathcal{C} \subseteq 2^X$ be a concept class. An **unlabelled sample compression scheme** or a USCS of size at most $k$ for $\mathcal{C}$ is a pair $(f, g)$ where $f : 2^{X \times \{0, 1\}} \to 2^X$, $g : 2^X \to 2^{X \times \{0, 1\}}$ such that if $S_C \subseteq X \times \{0, 1\}$ is a sample labelled consistently with some $C \in \mathcal{C}$, then

i) $f(S_C) \subseteq X(S_C)$,
ii) \( |f(S_C)| \leq k \),

iii) \( S \subseteq g(f(S_C)) \),

iv) \( g(f(S_C)) = S_{C'} \) for some \( C' \subseteq X \) and \( |S_{C'}| = |X| \).

We say that \((f, g)\) **has size** \( k' \) if \( k' \) is the smallest \( k \) for which \((f, g)\) is a unlabelled sample compression scheme of size at most \( k \). Here we call \( f(S_C) \) a **compression set**.

**Example 2.4.2.** Here we revisit the example of the axis-parallel rectangles, denote this set by \( R \subseteq \mathbb{R}^2 \) as before. Despite the fact that using unlabelled sets reduces the number of compression sets available, it is possible to construct a USCS of size 4 for \( R \). To do this we let \( R \in R \) be an axis-parallel rectangle and let \( S_R \subseteq 2^X \times \{0, 1\} \) be a sample labelled consistently with \( R \). Similar to the unlabelled compression scheme, if \( |\mathbf{1}(S_R)| \geq 4 \) the compression function \( f \) takes the bottommost, topmost, rightmost and leftmost positively labelled examples from \( S_R \), call them \((x_1, 1), (x_2, 1), (x_3, 1), (x_4, 1)\), and returns the set \( f(S_R) = \{x_1, x_2, x_3, x_4\} \) as the compression set for this sample. If \( |\mathbf{1}(S_R)| < 4 \) then \( f(S_R) = \mathbf{1}(S_R) \). The reconstruction function returns the smallest axis-parallel rectangle from \( R \) that has each point from \( f(S_R) \) labelled 1. In Example 2.3.2, the compression sets in the SCS for \( R \), each domain point is labelled with 1. Since we can consider a set of unlabelled points as a set of points all given the same labelling, then the fact that we have an SCS of size 4 for \( R \) implies that we have a
In [18], Kuzmin and Warmuth construct a USCS for maximum classes of size equal to the VC dimension. To do this, they define the notion of a representation mapping which maps concepts in a maximum concept class to subsets of the instance space of the class. This mapping ends up being a bijection since for a maximum class there are \( \binom{m}{\leq d} \) concepts while there are \( \binom{m}{\leq d} \) compression sets available under a USCS over a domain \( X \) with \( |X| = m \). Since the goal is not to compress the concepts themselves, but samples consistent with the concepts, the representation mapping must ensure that each of the representatives are spread out as much as possible with respect to each other to ensure that each sample has some representative that it can be mapped to.

**Definition 2.4.3.** [18] Let \( C \subseteq 2^X \) for some instance space \( X \) and let \( r : C \rightarrow \binom{X}{\leq \text{VCD}(C)} \). We say that two concepts \( C, C' \in C \) **clash** with respect to \( r \) if

\[
C|_{r(C) \cup r(C')} = C'|_{r(C) \cup r(C')}
\]

**Definition 2.4.4.** [18] A **representation mapping** \( r \) of a maximum concept class \( C \) is a function \( r : C \rightarrow \binom{X}{\leq \text{VCD}(C)} \) such that

a) \( r \) is a bijection,

b) no two concepts in \( C \) clash with respect to \( r \).
The no-clashing condition in Definition 2.4.4 allows the use of the representation mapping as the compression function for a USCS for a maximum class, and since it is a bijection, this function can be used to define the reconstruction function as well.

**Lemma 2.4.5.** [18] Let \( r : \mathcal{C} \rightarrow (\underline{\text{Dom}(\mathcal{C})}) \) be a bijection. Then the following statements are equivalent:

a) No two concepts clash with respect to \( r \).

b) For all samples \( S \subseteq X \times \{0, 1\} \) of a concept from \( \mathcal{C} \), there is exactly one concept \( C \in \mathcal{C} \) that is consistent with \( S \) and \( r(C) \subseteq X(S) \).

**Theorem 2.4.6.** [18] Let \( \mathcal{C} \subseteq 2^X \) be a maximum concept class over an instance space \( X \) and let \( r : \mathcal{C} \rightarrow (\underline{\text{Dom}(\mathcal{C})}) \) be a representation mapping for \( \mathcal{C} \). This mapping defines an unlabelled sample compression scheme for \( \mathcal{C} \) of size \( \text{VCD}(\mathcal{C}) \).

Thus to construct the unlabelled sample compression scheme for a maximum class \( \mathcal{C} \) of size equal to the VC dimension of \( \mathcal{C} \) it is sufficient to construct a representation mapping \( r : \mathcal{C} \rightarrow (\underline{X}) \). In [18] Kuzmin and Warmuth define an algorithm, called the tail matching algorithm, that produces such a representation mapping.

**Example 2.4.7.** An example of the output produced by the tail matching algorithm is given in [18]. Here we consider the maximum concept class \( \mathcal{C} \) of VC dimension 2 defined by the incidence matrix \( I_\mathcal{C} \). We also give the image of the representation mapping for each concept \( C \in \mathcal{C} \). This is expressed as the matrix \([I_\mathcal{C}|_{r(C)}]\) in Table 22.
2.4 on page 34. By Theorem 2.4.6, this representation mapping defines a USCS for \( \mathcal{C} \) of size 2.

We end this section with a simple observation that will be useful to help establish the significance of our results on unlabelled sequence-based sample compression.

**Lemma 2.4.8.** Let \( \mathcal{C} \subseteq 2^X \) be a maximum concept class of VC dimension \( d \). There is no unlabelled sample compression scheme of size less than \( d \) for \( \mathcal{C} \).

**Proof.** This lemma can be verified by simple counting. Any USCS for a maximum class \( \mathcal{C} \) of VC dimension \( d \) must at least distinguish each possible concept in \( \mathcal{C} \). However there are \( \binom{m}{\leq d} \) concepts in \( \mathcal{C} \) and only \( \binom{m}{\leq k} \) possible compression sets in any USCS of size \( k \) for \( \mathcal{C} \). So if \( k < d \) then \( \binom{m}{\leq k} < \binom{m}{\leq d} \). \( \square \)

### 2.5 Motivation and Conjectures

The primary motivation for studying sample compression schemes is for the implications they have for PAC-Learnability. To discuss the connections between these two topics we first define the notion of PAC-Learning and then discuss its relation to sample compression schemes.

In PAC learning, we begin with a concept \( C' \in \mathcal{C} \), where \( \mathcal{C} \) is a concept class, that we wish to learn. We will refer to this concept as the **target concept**. For learning to be meaningful, there must be some way to measure how well a concept is learned.
For a probability distribution $\mathcal{D}$, a target concept $C'$ and for a concept $C \in \mathcal{C}$ define the generalization error of $C$ to be the quantity

$$
err_{\mathcal{D},C'}(C) = P(\{x \in X \mid C'(x) \neq C(x)\}),
$$

for a concept $C \in \mathcal{C}$. Given the probability distribution $\mathcal{D}$ the generalization error of $C$, measures how much the concept returned by a learner for $C$ deviates from a target concept $C'$ that this learner wishes to learn. Ideally to PAC-learn some concept we will need to have a learner be able to draw enough examples to make the generalization error arbitrarily small.

**Definition 2.5.1.** [3][7] A concept class $\mathcal{C} \subseteq 2^X$ is said to be **PAC-learnable** if there exists an algorithm $L$ and a polynomial $q\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right)$ such that

- for all real numbers $\epsilon, \delta > 0$,
- for any probability distribution $\mathcal{D}$ over $X$, and
- for all $C' \in \mathcal{C}$ with $err_{\mathcal{D},C'}(C') = 0$ (ie. $C'$ is a target concept to be learned),

if $L$ draws a sample $S = \{(x_1, C'(x_1)), (x_2, C'(x_2)), \ldots, (x_n, C'(x_n))\} \subseteq X \times \{0, 1\}$ with $|S| = n \geq q\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right)$ and $x_1, x_2, \ldots, x_n$ are sampled independent and identically distributed with respect to $\mathcal{D}$, then $L$ returns a concept $C \in \mathcal{C}$ such that

$$
P(err_{\mathcal{D}}(C) \leq \epsilon) \geq 1 - \delta.
$$

We say that the **sample complexity** of $L$ on $\mathcal{C}$ is $n = n(\epsilon, \delta)$. 

24
Theorem 2.5.2. If \( \mathcal{C} \subseteq 2^X \) is a concept class, then there exists a learning algorithm \( L \) that PAC-learns with sample complexity \( \frac{1}{\epsilon} \ln \frac{|\mathcal{C}|}{\delta} \) for any \( \epsilon, \delta > 0 \).

In [7] Blumer et al. call a concept class well behaved if it satisfies certain measurability conditions. We omit the details of this definition as it would take us too far off the main topic of this thesis. Floyd and Warmuth note in [11] that this definition would be unlikely to exclude any concept classes that naturally arise in a learning theory context.

Theorem 2.5.3. If \( \mathcal{C} \subseteq 2^X \) is a well-behaved concept class, then there exists a learning algorithm \( L \) that PAC-learns with sample complexity

\[
\max \left( \frac{4}{\epsilon} \log \frac{2}{\delta}, \frac{8d}{\epsilon} \log \frac{13}{\epsilon} \right).
\]

for \( 0 < \epsilon, \delta < 1 \).

Notice above that the bounds for learning depend linearly on the VC dimension. This can be extended to show that the finiteness of the VC dimension provides a necessary and sufficient condition on the PAC-learnability of a concept class.

Theorem 2.5.4. [3][7] Let \( \mathcal{C} \subseteq 2^X \). The class \( \mathcal{C} \) is PAC learnable if and only if the VC dimension of \( \mathcal{C} \) is finite.

Theorem 2.5.5. [4][11] Let \( \mathcal{C} \subseteq 2^X \) be any concept class with a sample compression scheme of size at most \( d \). Then there exists a learning algorithm \( L \) such that for
$0 < \epsilon, \delta < 1$ $L$ PAC-learns $C$ with sample size

\[ m \geq \frac{1}{1 - \beta} \left( \frac{1}{\epsilon} \ln \frac{1}{\delta} + d + \frac{d}{\epsilon} \ln \frac{1}{\beta \epsilon} \right) \]

for any $0 < \beta < 1$.

The following conjecture motivates not only the work here, but more broadly, the study of sample compression schemes in general.

**Conjecture 2.5.6.** [11] Any concept class $C \subseteq 2^X$ with VC-dimension $d$ has a compression scheme of size $d$.

Resolving this conjecture positively would suggest that sample compression schemes not only offer an alternate approach to looking at PAC learning, but that it is intimately connected with the VC theoretical approach first proposed in [7]. The notion of compression to be introduced in this paper is somewhat more general than the original definition of sample compression schemes, however, resolving this conjecture for sequences-based compression schemes, would give evidence towards the positive resolution of Conjecture 2.5.6.

One of the challenges with resolving Conjecture 2.5.6 is that it allows for the possibility that $X$ could be infinite. Fortunately, in [13], Ben-David and Litman show that we need only be concerned with showing Conjecture 2.5.6 holds for all finite concept classes. Furthermore, though the next theorem is stated only for unlabelled sample compression schemes, it must hold for the labelled case since any unlabelled
sample compression scheme is a labelled compression scheme with fixed labellings of the points.

**Definition 2.5.7.** Let $\mathcal{C}$ be a concept class over an instance space $X$. If $Y \subseteq X$ and $\mathcal{C}' \subseteq 2^Y$, we call $\mathcal{C}'$ a **subclass** of $\mathcal{C}$ if $\mathcal{C}' = \mathcal{C}|_Y$.

**Theorem 2.5.8** (Compactness Theorem for Sample Compression Schemes). [13] A concept class $\mathcal{C}$ has an unlabelled sample compression scheme of size at most $d$ if and only if every finite subclass of $\mathcal{C}$ has an unlabelled sample compression scheme of size at most $d$.

It is because of the Compactness Theorem for sample compression schemes that we only need to show Conjecture 2.5.6 holds for finite concept classes. It is in this light that we restrict our analysis to finite concept classes in this work. Furthermore, Kalajdzievski noted in [25] that Theorem 2.5.8 may be extended to hold for a generalization of compression schemes. This generalization is called a $\{n_i\}_{i=0}^k$-copy sample compression schemes and, after we give some notation, we will formally define it.

For a finite sequence $\{n_i\}_{i=0}^k \subseteq \mathbb{N}$ let $(F, l) \in \bigcup_{i \in \{j \in \mathbb{N} : n_j \in \mathbb{Z}^+\}} (\binom{X}{i} \times [n_i])$. Define the **domain** of $(F, l)$ to be the set

$$\text{Dom}((F, l)) = F \subseteq X,$$

and the **size** of $(F, l)$ to be the non-negative integer

$$\| (F, l) \| = |F| \in \mathbb{N}.$$
Definition 2.5.9 ($\{n_i\}_{i=0}^k$-copy Unlabelled Sample Compression Schemes). Let $\{n_i\}_{i=0}^k \subseteq \mathbb{N}$ be a finite sequence, $X$ be an instance space and $\mathcal{C} \subseteq 2^X$ be a concept class. Further define

$$\mathcal{K} = \bigcup_{i \in \{j \in \mathbb{N} : n_j \in \mathbb{Z}^+\}} \left( \binom{X}{i} \times \{\underbrace{0, 1}_i\} \right).$$

An $\{n_i\}_{i=0}^k$-copy unlabelled sample compression scheme of size at most $k$ for $\mathcal{C}$ is a pair $(f, g)$ where $f : 2^X \times \{0, 1\} \rightarrow \mathcal{K}$, $g : \mathcal{K} \rightarrow 2^X \times \{0, 1\}$ such that if $S_C \subseteq X \times \{0, 1\}$ is a sample labelled consistently with some $C \in \mathcal{C}$, then

i) $\text{Dom}(f(S_C)) \subseteq X(S_C)$,

ii) $\|f(S_C)\| \leq k$,

iii) $S_C \subseteq g(f(S_C))$,

iv) $g(f(S_C)) = S_{C'}$ for some $C' \subseteq X$ and $|S_{C'}| = |X|$.

We say that $(f, g)$ has size $k'$ if $k'$ is the smallest $k$ for which $(f, g)$ is a $\{n_i\}_{i=0}^k$-copy unlabelled sample compression schemes of size at most $k$.

The intuition here is that we now have $n_i$ ‘copies’ of any compression set, labelled or unlabelled, of size $i$, and we keep track of which copy we are using by pairing the compression set with a positive integer.

Example 2.5.10. As an example, consider the standard maximum class of VC dimension 2. This is $\mathcal{C} = \mathcal{K}_2 \subseteq 2^X$, its incidence matrix is $I_\mathcal{C}$, given in Table 2.5 on page 35.
We can define a \( \{n_i\}_{i=0}^1 \)-copy sample compression scheme, \((f, g)\), of size 1 if we let \( n_0 = 1 \) and \( n_1 = 3 \). For a sample \( S \) consistent with \( C \in \mathcal{C} \) there are 3 possibilities, namely, \(|1(S)| = 0\), \(|1(S)| = 1\) and \(|1(S)| = 2\).

- If \(|1(S)| = 0\) we let \( f(S) = (\emptyset, 1) \).
- If \(|1(S)| = 1\) we let \( f(S) = (1(S), 1) \).
- If \(|1(S)| = 2\) then there is only one \( C \in \mathcal{C} \) such that \( S = S_C \). Thus we make the following assignment

\[
\begin{align*}
    f(S_C_6) &= (\{1\}, 2) \\
    f(S_C_7) &= (\{3\}, 2) \\
    f(S_C_8) &= (\{4\}, 2) \\
    f(S_C_9) &= (\{2\}, 2) \\
    f(S_C_{10}) &= (\{2\}, 3) \\
    f(S_C_{11}) &= (\{3\}, 3)
\end{align*}
\]

For the reconstruction function \( g \) we let \( (D, n) \in \mathcal{D} \) where \( \mathcal{D} \) is the set of compression sets created from applying \( f \) to the set of samples consistent with concepts in \( \mathcal{C} \), here \( n \) depends on the set \( D \).

- If \(|D| = 0\) then \( g((D, n)) = \{(x, 0) : x \in X\} \)
• If $|D| > 0$ and $n = 1$ then

\[ g((D, n)) = \{(x, y) : y = 1 \text{ if } x \in D \text{ and } y = 0 \text{ elsewhere}\} \]

• If $|D| > 0$ and $n = 2, 3$ then

\[ g((\{1\}, 2)) = \{(x, y) : y = 1 \text{ if } x \in C_6 \text{ and } y = 0 \text{ elsewhere}\} \]
\[ g((\{3\}, 2)) = \{(x, y) : y = 1 \text{ if } x \in C_7 \text{ and } y = 0 \text{ elsewhere}\} \]
\[ g((\{4\}, 2)) = \{(x, y) : y = 1 \text{ if } x \in C_8 \text{ and } y = 0 \text{ elsewhere}\} \]
\[ g((\{2\}, 2)) = \{(x, y) : y = 1 \text{ if } x \in C_9 \text{ and } y = 0 \text{ elsewhere}\} \]
\[ g((\{1\}, 3)) = \{(x, y) : y = 1 \text{ if } x \in C_8 \text{ and } y = 0 \text{ elsewhere}\} \]
\[ g((\{2\}, 3)) = \{(x, y) : y = 1 \text{ if } x \in C_{10} \text{ and } y = 0 \text{ elsewhere}\} \]
\[ g((\{3\}, 3)) = \{(x, y) : y = 1 \text{ if } x \in C_{11} \text{ and } y = 0 \text{ elsewhere}\} \]
\[ g((\{4\}, 3)) = \{(x, y) : y = 1 \text{ if } x \in C_{11} \text{ and } y = 0 \text{ elsewhere}\} \]

Notice here that for the pairs $(\{1\}, 3)$ and $(\{4\}, 3)$, even though they have a value that is returned, they do not play a role in the compression scheme since no compression function returns these pairs as a value. However, they ensure that $g$ is well defined.

Using these definitions of $f$ and $g$ it is clear that $(f, g)$ is a $\{n_i\}_{i=0}^1$-copy sample compression scheme of size 1. In terms of size, this offers some improvement over
sample compression schemes since for $K_2$ there cannot exist a sample compression scheme of size 1 since there would be at most $\binom{4}{0}2^0 + \binom{4}{1}2^1 = 9$ compression sets, while there are 11 concepts in $C$.

In similar fashion as for the original Compactness Theorem, we state the following theorem for unlabelled $\{n_i\}_{i=0}^k$-copy sample compression schemes and by extension this property must hold for labelled $\{n_i\}_{i=0}^k$-copy sample compression schemes as well.

**Theorem 2.5.11** (Compactness Theorem for $\{n_i\}_{i=0}^k$-copy sample compression schemes).

[25] Let $\{n_i\}_{i=0}^k \subseteq \mathbb{N}$ be a finite sequence and let $C \subseteq 2^X$ be a concept class. $C$ has an $\{n_i\}_{i=0}^k$-copy unlabelled sample compression scheme of size at most $d$ if and only if every finite subclass of $C$ has a unlabelled sample compression scheme of size at most $d$.

Based on the definition of a $\{n_i\}_{i=0}^k$-copy sample compression scheme a sequence-based compression scheme of size $k$ is just a $\{i!\}_{i=0}^k$-copy compression scheme of size $k$ so for sequence-based compression, just like with standard sample compression schemes, we only need to be concerned with compressing finite concept classes.

**Theorem 2.5.12.** [25] Let $C \subseteq 2^X$ be a concept class and let $\{n_i\}_{i=0}^k \subseteq \mathbb{N}$ be a finite sequence. If $C$ has a $\{n_i\}_{i=0}^k$-copy sample compression schemes of size $k$ and if $l = \max(\{n_i\}_{i=0}^k)$ then for $0 < \epsilon, \delta \leq 1$, there exists a learning algorithm $L$ that
PAC-learns $C$ with sample size

$$m \geq \frac{1}{1 - \beta} \left( \frac{1}{\epsilon} \ln \frac{1}{\delta} + k + \frac{k}{\epsilon} \ln \frac{1}{\beta \epsilon} \right)$$

for any $0 < \beta < 1$.

Therefore, since sequence-based sample compression schemes, both labelled and unlabelled, are just $\{i!\}_{i=0}^{k}$-copy sample compression schemes, then sequence-based sample compression implies learnability.

**Corollary 2.5.13.** Let $C \subseteq 2^X$ be a concept class. If $C$ has a unlabelled sequence-based sample compression scheme of size $k$ then there exists a learning algorithm $L$ that PAC-learns $C$ with sample size

$$m \geq \frac{1}{1 - \beta} \left( \frac{1}{\epsilon} \ln \frac{k!}{\delta} + k + \frac{k}{\epsilon} \ln \frac{1}{\beta \epsilon} \right)$$

for any $0 < \beta < 1$.

The downside of obtaining this result via $\{n_i\}_{i=0}^{k}$-copy sample compression schemes is that for sequence-based sample compression schemes, $\max(\{n_i\}_{i=0}^{k}) = \max(\{i!\}_{i=0}^{k}) = k!$. Since the sample complexity in Theorem 2.5.12 is linear in $\max(\{n_i\}_{i=0}^{k})$ it does not show any improvement in sample complexity over regular sample compression schemes.
<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$C_3$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_4$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_5$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_6$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$C_7$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_8$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$C_9$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_{10}$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_{11}$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$C_{12}$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_{13}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_{14}$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$C_{15}$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$C_{16}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2.3: Maximum concept class of VC dimension 2
<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$C_2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>${x_3}$</td>
</tr>
<tr>
<td>$C_3$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>${x_4}$</td>
</tr>
<tr>
<td>$C_4$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>${x_2}$</td>
</tr>
<tr>
<td>$C_5$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>${x_1}$</td>
</tr>
<tr>
<td>$C_6$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>${x_1, x_3}$</td>
</tr>
<tr>
<td>$C_7$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>${x_1, x_4}$</td>
</tr>
<tr>
<td>$C_8$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>${x_1, x_2}$</td>
</tr>
<tr>
<td>$C_9$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>${x_3, x_4}$</td>
</tr>
<tr>
<td>$C_{10}$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>${x_2, x_3}$</td>
</tr>
<tr>
<td>$C_{11}$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>${x_2, x_4}$</td>
</tr>
</tbody>
</table>

Table 2.4: Representation mappings for a maximum class of VC dimension 2
Table 2.5: Standard Maximum Class of VC dimension 2

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_2$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_3$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_4$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$C_5$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$C_6$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_7$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$C_8$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$C_9$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$C_{10}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$C_{11}$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Chapter 3

A Framework for Defining New Sample Compression Schemes

For defining sequence-based sample compression schemes, first we will generalize the concept of compression schemes, then use this generalization to define sequence-based sample compression.

3.1 Generalized Notion of Sample Compression Schemes

Definition 3.1.1 (Compression Schemes). Let $X, B$ and $\mathcal{K}$ be sets and let $\mathcal{C} \subseteq 2^X$ be a concept class. Let $\Phi_1 : \mathcal{K} \to B$ and $\Phi_2 : 2^{X \times \{0,1\}} \to B$ be functions. Finally let $\text{SIZE} : \mathcal{K} \to \mathbb{N}$ and let $k \in \mathbb{N}$. A $(\Phi_1, \Phi_2, \text{SIZE}, \mathcal{K}, B)$-compression scheme of size at most $k$ for $\mathcal{C}$ is a pair $(f, g)$ where $f : 2^{X \times \{0,1\}} \to \mathcal{K}$, $g : \mathcal{K} \to 2^{X \times \{0,1\}}$ such that if $S_C \subseteq X \times \{0,1\}$ is a sample labelled consistently with some $C \in \mathcal{C}$, then
i) \( \Phi_1(f(S_C)) \subseteq \Phi_2(S_C) \),

ii) \( \text{SIZE}(f(S_C)) \leq k \),

iii) \( S \subseteq g(f(S_C)) \),

iv) \( g(f(S_C)) = S_{C'} \) for some \( C' \subseteq X \) and \( |S_{C'}| = |X| \).

We say that \( (f, g) \) has size \( k' \) if \( k' \) is the smallest \( k \) for which \( (f, g) \) is a \((\Phi_1, \Phi_2, \text{SIZE}, \mathcal{K}, \mathcal{B})\)-compression scheme of size at most \( k \).

The functions \( \Phi_1 \) and \( \Phi_2 \) here allow for defining more general inclusions of sets. For instance, in Definition 2.3.1 we require for a sample \( S_C \) consistent with some concept \( C \in \mathcal{C} \) that \( f(S_C) \subseteq S_C \) where \( f \) is a compression function for a \( \mathcal{SCS} \). On the other hand in Definition 2.4.1 we must require that \( f'(S_C) \subseteq X(S_C) \) where \( f' \) is the compression function for a \( \mathcal{USCS} \). Thus to capture condition (i) in each of these definitions we must have some way to refer to some subset or subsequence of a sample \( S_C \). Also notice that in Definition 3.1.1, for our compression functions the domain of \( g \) and codomain of \( f \) are fixed by our definitions of \( \Phi_1 \) and \( \Phi_2 \).

Let \( \mathcal{C} \subseteq 2^X \) and let \( \text{id}_{X \times \{0,1\}} : 2^{X \times \{0,1\}} \to 2^{X \times \{0,1\}} \) be the function defined by

\[
\text{id}_{X \times \{0,1\}}(S) = S
\]

for all \( S \in 2^{X \times \{0,1\}} \). A sample compression scheme of size \( k \in \mathbb{N} \) for \( \mathcal{C} \) is a \((\text{id}_{X \times \{0,1\}}, \text{id}_{X \times \{0,1\}}, |\cdot|, 2^{X \times \{0,1\}}, 2^{X \times \{0,1\}})\)-compression scheme \((f, g)\) of size \( k \in \mathbb{N} \) for \( \mathcal{C} \).
Furthermore we can also use this definition to define the unlabelled sample compression schemes from [18]. Let $C \subseteq 2^X$ and let $id : 2^{X \times \{0,1\}} \to 2^{X \times \{0,1\}}$, $id_X : 2^{X \times \{0,1\}} \to 2^X$ be the functions defined by

$$id_{X \times \{0,1\}}(S) = S, \quad id_X(S) = X(S)$$

for all $S \in 2^{X \times \{0,1\}}$. Then an unlabelled sample compression scheme of size $k \in \mathbb{N}$ for $C$ is a $(id_X \times \{0,1\}, id_X, | \cdot |, 2^X, 2^X)$-compression scheme $(f, g)$ of size $k \in \mathbb{N}$ for $C$.

### 3.2 Sequence-Based Compression

The general definition of compression schemes can be further specified to describe sequence-based compression schemes.

#### 3.2.1 Definitions and Notation

In order to define sequence based compression, we must first define the language and notation which we will use to discuss this topic.

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set with a total ordering given by $x_i < x_j$ if $i < j$. We denote the set of sequences of points taken from $X$ by $X^*$. For a sequence $(D) = (x_{i_1}, x_{i_2}, \ldots, x_{i_r}) \in X^*$ where $i_j \leq i_{j+1}$ for all $j \in \{1, 2, \ldots, r-1\}$ we say that $(D)$ has the **standard ordering** with respect to $X$ and will denote this $(D)_{stan}$. If $(x_{i_1}, x_{i_2}, \ldots, x_{i_r})$ where $i_j \geq i_{j+1}$ for all $j \in \{1, 2, \ldots, r-1\}$, then we say that $(D)$ has the **reverse ordering** with respect to $X$ and denote this $(D)_{rev}$. 

38
Next if \( (A) = (a_1, a_2, \ldots, a_k) \) then we say the length of \((A)\) is \(k\) and write \(|(A)| = k\), and we let \(\Lambda\) denote the **empty sequence**, that is, the sequence with length 0.

**Definition 3.2.1.** Let \((A) = (a_1, a_2, \ldots, a_k)\) be any finite sequence of elements from a set \(A\). If \(a\) occurs in the sequence \((A)\) we write \(a \prec (A)\). Define the **domain** of the sequence to be the set

\[
\text{Dom}(A) = \{ a \in A : a \prec (A) \}.
\]

For a set \(B\) if \(\text{Dom}((A)) \subseteq B\) we write \((A) \subseteq B\) and call \((A)\) a subsequence of \(B\).

We call any \((S) \in (X \times \{0,1\})^*\) a **labelled sequence** and define the **instance set** of \(S^*\) as the set

\[
X^*((S)) = \{ x \in X : (x, y) \in \text{Dom}((S)) \text{ for some } y \in \{0,1\} \}.
\]

Finally if \(C\) is a concept class we let \(C^* = \{(C) \in X^* : C \in C\}\).

Thus using this notation, we have

\[
\binom{A}{d}^* := \{ \omega \in A^* : |\omega| = d \}
\]

and

\[
\binom{A}{\leq d}^* := \{ \omega \in A^* : |\omega| \leq d \}.
\]

Although for a set \(X\) and a concept class \(C\) the notation \(X^*\) and \(C^*\) differ with respect to each other, we do not consider sequences of sets in this work so it will be clear which set of sequences we are referring to.
3.2.2 Sequence-Based Compression Schemes

**Definition 3.2.2.** [Labelled Sequence-Based Sample Compression Scheme] Let $\mathcal{C} \subseteq 2^X$ and let $\theta : (X \times \{0, 1\})^* \rightarrow 2^{X \times \{0, 1\}}$, $id_{X \times \{0, 1\}} : 2^{X \times \{0, 1\}} \rightarrow 2^{X \times \{0, 1\}}$ be the functions defined by
\[
\theta((D)) = \text{Dom}(D)
\]
for $D \in 2^{X \times \{0, 1\}}$, and
\[
id_{X \times \{0, 1\}}(S) = S
\]
for all $S \in 2^{X \times \{0, 1\}}$. A **Sequence-Based Sample Compression Scheme**, abbreviated $\text{SBSCS}$, of size $k \in \mathbb{N}$ for $\mathcal{C}$ is a $(\theta, id_{X \times \{0, 1\}}, | \cdot |, (X \times \{0, 1\})^*, 2^{X \times \{0, 1\}})$-compression scheme $(f^*, g^*)$ of size $k \in \mathbb{N}$ for $\mathcal{C}$. Here we call $f(S_C)$ a compression sequence.

**Definition 3.2.3.** [Unlabelled Sequence-Based Sample Compression Scheme] Let $\mathcal{C} \subseteq 2^X$ and let $\sigma : X^* \rightarrow 2^X$, $id_X : 2^{X \times \{0, 1\}} \rightarrow 2^X$ be the functions defined by
\[
\sigma((D)) = \text{Dom}(D)
\]
for all $D \in 2^X$, and
\[
id_X(S) = X(S)
\]
for all $S \in 2^{X \times \{0, 1\}}$. An **Unlabelled Sequence-Based Sample Compression Scheme**, abbreviated $\text{USBSCS}$, of size $k \in \mathbb{N}$ for $\mathcal{C}$ is a $(\sigma, id_X, | \cdot |, X^*, 2^X)$-compression scheme $(f^*, g^*)$ of size $k \in \mathbb{N}$ for $\mathcal{C}$. Here we call $f(S_C)$ a compression sequence.
sequence.

3.2.3 Relation to \( \{n_i\}_{i=0}^k \)-copy sample compression schemes

This generalized definition of compression schemes is also capable of defining other generalizations of compression schemes. Namely, the \( \{n_i\}_{i=0}^k \)-copy sample compression scheme defined by Kalajdzievski in [25] can be expressed under this framework.

Now let \( \{n_i\}_{i=0}^k \subseteq \mathbb{N} \) be a finite sequence and let

\[
\mathcal{K} = \bigcup_{i \in \{j \in \mathbb{N} : n_j \in \mathbb{Z}^+\}} \left( \left( \frac{X}{i} \right) \times [n_i] \right).
\]

We use an alternate notation for finite sequences here to prevent confusion with the sequences used for compression in sequence-based compression schemes.

Let \( \mathcal{C} \subseteq 2^X \) and let \( \gamma : \mathcal{K} \rightarrow 2^X, id_X : 2^X \rightarrow 2^X \) be the functions defined by

\[
\gamma(D) = \text{Dom}(D),
\]

\[
id_X(S) = X(S)
\]

for all \( S \in 2^{X \times \{0,1\}}, D \in \mathcal{K} \). Then an \( \{n_i\}_{i=0}^k \)-copy Unlabelled Sample Compression scheme of size \( k \in \mathbb{N} \) for \( \mathcal{C} \) is a \((\gamma, id_X, \| \cdot \|, \mathcal{K}, 2^X)\)-compression scheme \((f, g)\) of size \( k \in \mathbb{N} \) for \( \mathcal{C} \).

Similarly, an \( \{n_i\}_{i=0}^k \)-copy labelled sample compression scheme of size \( k \in \mathbb{N} \) for \( \mathcal{C} \) is a \((\rho, id_{X \times \{0,1\}}, \| \cdot \|, \mathcal{K'}, 2^{X \times \{0,1\}})\)-compression scheme of size \( k \in \mathbb{N} \) for \( \mathcal{C} \) where

\[
\mathcal{K'} = \bigcup_{i \in \{j \in \mathbb{N} : n_j \in \mathbb{Z}^+\}} \left( \left( \frac{X \times \{0,1\}}{i} \right) \times [n_i] \right).
\]
and

\[ \rho(D) = \text{Dom}(D), \]
\[ id_{X \times \{0,1\}}(S) = S \]

for all \( S \in 2^{X \times \{0,1\}}, D \in \mathcal{K}' \).

Based on the definition we have used for the domain, for \( D \in \mathcal{K} \) and \( D' \in \mathcal{K}' \) \( \gamma(D) = \text{Dom}(D) \) is an unlabelled set while \( \rho(D) = \text{Dom}(D) \) is an unlabelled set.
Chapter 4

Compression Results

4.1 Unlabelled Sequence-Based Sample Compression Schemes

4.1.1 The Size $\left\lfloor \frac{m}{2} \right\rfloor$ unlabelled Sequence-Based Sample Compression Scheme Algorithm

The following scheme produces a unlabelled sequence-based sample compression scheme of size $\left\lfloor \frac{m}{2} \right\rfloor$ for a concept class $C$ over $X$ with $|X| = m \geq 6$. The input to our compression scheme is a sample set $S \subseteq 2^{X \times \{0,1\}}$ that is consistent with a concept in $C$. To describe the algorithm for compression and reconstruction, we will specify how the function pair $(f^*, g^*)$ behaves given a specific domain input.

The $\left\lfloor \frac{m}{2} \right\rfloor$ unlabelled sequence-based sample compression scheme algorithm

The Compression Function $f^*$: Defining the compression function $f^*$ requires
separating a number of disjoint cases.

**Case 1:** If \(|1(S)| = 0\), then \(f^*(S) = \Lambda\).

**Case 2:** If \(0 < |1(S)| \leq \lfloor \frac{m}{2} \rfloor\), then \(f^*(S) = (1(S))_{\text{stan}}\).

**Case 3:** If \(|1(S)| > \lfloor \frac{m}{2} \rfloor\), then we must consider a few subcases that arise.

1. If \(|0(S)| > 1\), then \(f^*(S) = (0(S))_{\text{rev}}\).
2. If \(|S| \geq 3\) and \(|0(S)| = 0\), then \(f^*(S) = (a, c, b)\) where \(a, b, c \in X(S) = 1(S)\) with \(a < b < c\).
3. If \(|S| \geq 3\) and \(|0(S)| = 1\), then for the point \((x, 0) \in S\) we have
   
   i) If \(x \leq x'\) for all \(x' \in X(S)\), then \(f^*(S) = (a, x, b)\) where \(a, b \in 1(S)\) with \(x < a < b\).
   
   ii) If \(x \geq x'\) for all \(x' \in X(S)\), then \(f^*(S) = (b, x, a)\) where \(a, b \in 1(S)\) with \(a < b < x\).
   
   iii) Otherwise, we let \(f^*(S) = (b, a, x)\) where \(a, b \in 1(S)\) with \(a < x < b\).

For \(m > 4\) we do not have to consider the case \(|1(S)| > \lfloor \frac{m}{2} \rfloor\) and \(|S| < 3\). This is because if \(|S| < 3\) then either \(|0(S)| = 2\), \(|0(S)| = 1\) or \(|0(S)| = 0\). In each of these cases, we have \(|1(S)| \leq 2 = \lfloor \frac{m}{2} \rfloor\) and so this case is covered by Case 2 in the algorithm.

**The Reconstruction Function** \(g^*\): Let \(D\) denote a compression sequence that is an input to our reconstruction function.
1. If $D = \Lambda$, then $g^*(D) = \{(x,0) \mid x \in X\}$.

2. If the input is a sequence $D = (x_1, \ldots, x_r)$ under the standard ordering, then

$$g^*(D) = \{(x_i,y) \in X \times \{0, 1\} \mid y = 1 \text{ if } x_i \in D \text{ and } y = 0 \text{ elsewhere}\}.$$ 

3. If the input is a sequence $D = (x_r, \ldots, x_1)$, $r > 2$, under the reverse ordering, then

$$g^*(D) = \{(x_i,y) \in X \times \{0, 1\} \mid y = 0 \text{ if } x_i \in D \text{ and } y = 1 \text{ elsewhere}\}.$$ 

4. Otherwise, assume $a, b, c \in X$ are the elements occurring in $D$ and $a < b < c$.

**Case 1:** If $D = (b,a,c)$, then

$$g^*(D) = \{(x,y) \mid y = 0 \text{ if } x = a, \text{ and } y = 1 \text{ elsewhere}\}.$$ 

**Case 2:** If $D = (b,c,a)$, then

$$g^*(D) = \{(x,y) \mid y = 0 \text{ if } x = c, \text{ and } y = 1 \text{ elsewhere}\}.$$ 

**Case 3:** If the input is any remaining ordering of size 3, then

$$g^*(D) = \{(x,y) \mid y = 0 \text{ if } x = b, \text{ and } y = 1 \text{ elsewhere}\}.$$ 

Despite the apparent complexity of this compression algorithm, intuitively, it is just making use of the fact that any sample set $S \subseteq 2^{X \times \{0,1\}}$ labelled consistently with a concept $C \in \mathcal{C}$ has either at most $\left\lfloor \frac{m}{2} \right\rfloor$ points labelled 1 or at most $\left\lceil \frac{m}{2} \right\rceil$ points labelled 0.
Theorem 4.1.1. The $\left\lfloor \frac{m}{2} \right\rfloor$ Unlabelled Sequence Scheme Algorithm is an Unlabelled Sequence-Based Sample Compression Scheme of size $\left\lfloor \frac{m}{2} \right\rfloor$ for any concept class $C \subseteq 2^X$ with $|X| = m > 5$ or $m = 4$.

Proof. Let $|X| = m \in \{4, 6, 7, 8, \ldots \}$ and let $S_C$ be a sample labelled consistently with some $C \in C$. To show that this algorithm performs correctly, we will show that each of the conditions that allow this algorithm to be a USBSCS of size $\left\lfloor \frac{m}{2} \right\rfloor$ for $C$ are satisfied.

1. For all $S \subseteq 2^{X \times \{0,1\}}$ we have that $\text{Dom}(f^*(S)) \subseteq X(S)$.

2. It is clear that in each case in the algorithm we have $|f^*(S)| \leq \left\lfloor \frac{m}{2} \right\rfloor$.

3. Here we need to show that $S_C \subseteq g^*(f^*(S_C))$. If $|\mathbf{1}(S)| = 0$ then

$$S_C \subseteq \{(x, 0) \mid x \in X\} = g^*(\Lambda) = g^*(f^*(S_C))$$

and if $0 < |\mathbf{1}(S)| \leq \left\lfloor \frac{m}{2} \right\rfloor$ then

$$S_C \subseteq g^*(\mathbf{1}(S_C)) = g^*(f^*(S_C)).$$

Now, if $|\mathbf{1}(S)| > \left\lfloor \frac{m}{2} \right\rfloor$, we notice in any case listed in our algorithm that either $f^*(S_C)$ is a sequence in the reverse ordering of length $r$ where $r \geq 3$ or it is some sequence of size 3 under some other ordering. The first case occurs when $|\mathbf{0}(S_C)| > 1$ and then $g^*(f^*(S_C)) = \{(x_i, y) \in X \times \{0, 1\} \mid y = 0 \text{ if } x_i \in f^*(S_C) \text{ and } y = 1 \text{ elsewhere}\}$, since $\text{Dom}(f^*(S_C)) \subseteq X(S_C)$ then
\( S_C \subseteq g^*(f^*(S_C)) \). The second case occurs when \(|0(S_C)| = 0\) or \(|0(S_C)| = 1\), in either case we have that \( S_C \subseteq g^*(0(S_C)) \).

4. In each case, for any compression set \( D \) there is no \( x \in \text{Dom}(g^*(D)) \) such that both \((x, 1), (x, 0) \in g^*(D)\) so that there exists some \( C' \subseteq X \) such that \( g^*(D) = S_{C'} \) and \(|S_{C'}| = |X|\) holds by definition.

\[ \square \]

**Corollary 4.1.2.** There exists a concept class \( C \subseteq 2^X \) such that the \( \lfloor \frac{m}{2} \rfloor \) unlabelled sequence-based sample compression scheme algorithm has size less than the VC dimension of \( C \).

**Proof.** Let \( K_d \subseteq 2^X \) be the standard maximum class over \( X \) where \( d = \text{VCD}(K_d) > \frac{m}{2} \) and \(|X| = m\). It is clear that the \( \lfloor \frac{m}{2} \rfloor \) unlabelled sequence-based sample compression scheme algorithm produces a USBS of size \( \frac{m}{2} < \text{VCD}(K_d) \) for \( K_d \). \[ \square \]

The previous corollary demonstrates the value of approaching compression schemes using sequences instead of sets. Specifically if we were to use unlabelled sets as the compression sets, since \( K_d \) is a maximum class of VC dimension \( d \) then by Lemma 2.4.8 \( K_d \) cannot have a USCS of size less than \( d \).

**Proposition 4.1.3.** If \(|X| \in \{2, 3, 5\}\), then there exists a \( C \subseteq 2^X \) such that there is no Unlabelled Sequence-Based Sample Compression scheme of size \( \lfloor \frac{m}{2} \rfloor \) for \( C \).
Proof. For $|X| = m \in \{2, 3, 5\}$ let $C = 2^X$. To compress $C$ we must have at least $2^m$ compression sequences, since we must at least be able to compress all concepts in our concept class. However, if $m \in \{2, 3, 5\}$, the number of compression sequences is

$$\sum_{i=0}^{\left\lfloor \frac{m}{2} \right\rfloor} \binom{m}{i} i! < 2^m.$$ 

Hence we do not have enough compression sequences to compress each of the $2^m$ concepts in $C$ and so an $SBSCS$ cannot exist for the concept class $C$. 

4.1.2 Unlabelled Sequence-Based Compression Scheme of Size Less than the VC Dimension

Using sequences it is possible to obtain a compression bound that is smaller than the VC dimension. To define this compression scheme, we first need to establish a result on matchings in a bipartite graph. This will allow us to define an injection from the subsets of $X$ with size equal to the VC dimension $d$, to subsequences of $X$ of size $d - 1$. This injection is apparent due to the fact that we can view the injection as a maximum matching on a bipartite graph. For this fact to help us, we will first need to recall some facts about bipartite graphs.

Definition 4.1.4. A bipartite graph is a graph $G = (V, E)$ such that we may partition the vertex set $V$ into two disjoint subsets $U$ and $W$ such that if $\{u, w\} \in E$ then $u \in U$ and $w \in W$. We refer to $U$ and $W$ as the bipartition sets.
Definition 4.1.5. Let $G = (V, E)$ be a graph, we call a subset $M \subseteq E$ a **matching** if $M \neq \emptyset$ and for all $M, N \in M$ we have that $M \cap N = \emptyset$.

Intuitively, a matching is just any set $F$ of edges from a graph such that no two edges in $F$ are incident to one another.

Definition 4.1.6. Let $G = (V, E)$ be a graph with vertex set $V$ and edge set $E$. We say that $v, w \in V$ are **neighbors** if $\{v, w\} \in E$. The set $N(v)$ of neighbors of $v \in V$ is called the **neighborhood** of $v$. Further if $A \subseteq V$, then we call the set $N(A) = \cup_{v \in A} N(v)$ the **neighbourhood** of $A$.

Theorem 4.1.7. [2] Let $G = (V, E)$ be a bipartite graph with bipartitions $U$ and $W$ with $r = |U| \leq |W|$. $G$ has a matching of cardinality $r$ if and only if $|N(A)| \geq |A|$ for all $A \subseteq U$.

Before our next lemma, we first define $D^* = \{(A) \in (\leq_{d-1})^* | (A) \neq (A)_{stan}\}$. We will use this notation to denote the set $\{(A) \in (\leq_{d-1})^* | (A) \neq (A)_{stan}\}$ for the rest of this section. So intuitively, this set $D^*$ is the set of all sequences of size up to $d - 1$ that are not in the standard ordering. Notice that $|D^*| = \sum_{i=2}^{d-1} \binom{m}{i}(i! - 1)$ and there are no sequences of size less than 2 in the set $D^*$.

Lemma 4.1.8. If $m \geq d \geq 3$ and $\binom{m}{d} \leq \sum_{i=2}^{d-1} \binom{m}{i}(i! - 1)$, then there exists an injection $\phi : (\leq_{d}) \rightarrow D^*$.
Proof. First define a bipartite graph \( G = (V, E) \) where \( V = M \cup K \), where \( M = (X_d) \) and \( K = D^* \), and \( E = \{(A, (B))|(B) \subseteq A, A \in M, (B) \in K\} \). Since \( \binom{m}{d} \leq D^* \) then \( |M| \leq |K| \) so if we could construct a matching of size \(|M|\) then this would correspond to an injection from the set \( M \) to the set \( K \). Using Theorem 4.1.7 we can say that such a matching exists if and only if \(|N(A)| \geq |A|\) for all \( A \subseteq M \). Let \( A \subseteq M \) and let \( a_1, \ldots, a_{|A|} \in A \). Since each \( a_i \) is distinct and \( d \geq 3 \) then for each \( a_i \) there exists some \( (b_i) \) with \(|(b)| = d - 1 \) such that \( (b_i) \not\subseteq a_j \) for any \( j \neq i \). This means that \(|A| \leq |N(A)|\) and since \( A \subseteq M \) was arbitrary then \(|N(A)| \geq |A|\) for all \( A \subseteq M \) so a matching of size \(|M|\) must exist.

For a fixed \( d \) as \( m \) grows larger we of course see that \( \binom{m}{d} \geq \sum_{i=2}^{d-1} \binom{m}{i}(i! - 1) \). However even for \( m \) as large as \( d! - 1 \) the bound \( \binom{m}{d} \leq \sum_{i=2}^{d-1} \binom{m}{i}(i! - 1) \) still holds.

To construct our unlabelled sequence-based sample compression scheme of size \( d - 1 \) we will first need to create an injection between the subsets of \( X \) of size \( d \) and the subsequences of \( X \) of size \( d - 1 \) that are not under the standard ordering.

**The \( d - 1 \) unlabelled sequence-based sample compression algorithm**

**The Compression function \( f^* \):** Here we let \( f^* \) be our compression function and \( \mathcal{C} \subseteq 2^X \) be a concept class with \( \text{VCD}(\mathcal{C}) = d \). Let \( \phi : (X_d) \to D^* \) be as predicted in the previous lemma and let \( S \subseteq 2^{X \times Y} \) be a sample consistent with \( \mathcal{C} \). Assume that we have a \( \mathcal{USCS} \), say \((f, g)\) of size \( d \) for \( \mathcal{C} \).
Case 1: If $f(S) \leq d - 1$, then $f^*(S) = (f(S))_{\text{stan}}$.

Case 2: If $f(S) = d$, then we let $f^*(S) = (\phi \circ f)(S)$.

The Reconstruction function $g^*$: Let $D$ denote a compression sequence that is input to our reconstruction function. Since $\phi$ is an injective function it follows that its range, $\phi((X^d))$ induces a bijection $\sigma$ that gives rise to an inverse map $\sigma^{-1}$:

$$\phi((X^d)) \rightarrow (X^d)$$

1. If $|D| < d - 1$, then $g^*(D) = g(\text{Dom}(D))$.

2. If $|D| = d - 1$, then we consider two cases.

Case 1: If $D$ has the standard ordering, then $g^*(D) = g(\text{Dom}(D))$.

Case 2: If $D$ does not have the standard ordering, then $g^*(D) = (g \circ \sigma^{-1})(D)$.

Theorem 4.1.9. Let $C \subseteq 2^X$ be a concept class over a finite set $X$ with a USCS of size $d$. If $m \geq d \geq 3$ and $\binom{m}{d} \leq \sum_{i=2}^{d-1} \binom{m}{i}(i! - 1)$ then $C$ has an unlabelled sequence-based sample compression scheme of size $d - 1$.

Proof. Let $S_C$ be a sample consistent with a concept in $C \in C$ where $C$ is a concept class of VC dimension $d$ and assume that $C$ has a USCS of size $d$ for $C$, denote this $(f, g)$. Further let $\phi$ and $\sigma^{-1}$ be the injections predicted in Lemma 4.1.8. Notice for any $D \in (X^d)$ we have that $\phi(\sigma^{-1}(D)) = \sigma^{-1}(\phi(D)) = D$. Let $(f^*, g^*)$ be the functions defined by the $d - 1$ unlabelled sequence-based sample compression algorithm. To
show that our compression function is a USBSCS of size $d - 1$ for $\mathcal{C}$ we proceed by verifying that each of the properties (i) to (iv) in the definition of a USBSCS are satisfied.

i) In both cases listed in the compression function $f^*$ we have that either

\[
\text{Dom}(f^*(S_C)) = (f(S_C))_{\text{stan}} \subseteq X(S_C)
\]

or

\[
\text{Dom}(f^*(S_C)) = (\phi \circ f)(S_C) \subseteq X(S_C).
\]

ii) Each sequence in the image of $f^*$ has size $d - 1$. Therefore $|f^*(S_C)| \leq d - 1$.

iii) To show that $S_C \subseteq g^*(f^*(S_C))$ we notice two cases. In the first case $f^*(S_C) = (f(S_C))_{\text{stan}}$. In this case our reconstruction function is given by

\[
g^*(f(S_C)) = g(\text{Dom}(f(S_C))) = g(f(S_C))
\]

and since $(f, g)$ is a USCS for $\mathcal{C}$ we have that $S_C \subseteq g(f(S_C))$. In the second case $f^*(S_C) = (\phi \circ f)(S_C)$. Now the sequence $f^*(S_C) = (\phi \circ f)(S_C)$ has either the standard ordering or it does not. If it has the standard ordering, then $S_C \subseteq g(f(S_C)) = g^*(f^*(S_C))$. If $f^*(S_C)$ does not have the standard ordering then
\[ g^*(f^*(S_C)) = (g \circ \sigma^{-1})((\phi \circ f)(S_C)) \]
\[ = g((\sigma^{-1} \circ \phi)(f(S_C))) \]
\[ = g(f(S_C)). \]

Thus \( S_C \subseteq g^*(f^*(S_C)) \).

iv) Since for any set \( D \in f^*(2^X) \) we have
\[ g^*(D) = g(\text{Dom}(D)) \]

or
\[ g^*(D) = (g \circ \sigma^{-1})(D) \]

and since \( g \) is a reconstruction function for a \( USCS \) for \( C \) then we must have that \( g^* \) satisfies condition (iv) of the definition of a \( USCS \) which is identical to condition (iv) in the definition of a \( USBSCS \).

\[ \square \]

Theoretically, the previous theorem shows that a \( USBSCS \) of size \( d - 1 \) is possible for any class \( C \) with a \( USCS \) of size \( d \). However, the details of this construction depend on the construction of the injection \( \phi \), which corresponds to the construction of a matching of a large enough size on a bipartite graph. Fortunately, the construction of such a matching is a well known. Using the Hopcroft-Karp algorithm [20] a maximum
matching can be constructed with $O(\sqrt{|V||E|})$ time complexity for any bipartite graph $G = (V, E)$. Since we have established that a matching of size $\binom{m}{d}$ exists on the graph $G$ constructed in Theorem 4.1.8, we know that such a maximum matching would have size greater than this, and would exactly meet the requirements necessary to make the matching correspond to an injection $\phi$.

**Corollary 4.1.10.** Let $\mathcal{C} \subseteq 2^X$ be a maximum concept class with $VCD(\mathcal{C}) = d$ over a finite set $X$. If $m \geq d \geq 3$ and $\binom{m}{d} \leq \sum_{i=2}^{d-1} \binom{m}{i}(i! - 1)$ then $\mathcal{C}$ has an unlabelled sequence-based sample compression scheme of size $d - 1$.

**Proof.** By [18], if $\mathcal{C} \subseteq 2^X \times \{0,1\}$ is a maximum class of VC dimension $d$ we know there exists an Unlabelled Sample Compression Scheme for $\mathcal{C}$ of size $d$. By Theorem 4.1.9, $\mathcal{C}$ must have an unlabelled sequence-based sample compression scheme of size $d - 1$. $\square$

This corollary, together with Lemma 2.4.8 demonstrates the usefulness of unlabelled sequences as opposed to labelled sets. In Lemma 2.4.8 we can see that for a maximum class of VC dimension $d$ there cannot exist a USCS of size less than $d$.

### 4.2 Labelled Sequence-Based Sample Compression Schemes

This section explores some of the properties and limitations of using labelled sequence-based sample compression schemes. The following theorem shows that any labelled sample compression scheme is also a labelled sequence-based sample compression scheme.
Theorem 4.2.1. Let $C \subseteq 2^X$ be a finite concept class over a set $X$ and let $k \in \mathbb{N}$.

There exists a labelled sample compression scheme for $C$ of size $k$ if and only if there exists a labelled sequence-based sample compression scheme $(f^*, g^*)$ for $C$ of size $k$ with the property that for any pair of distinct samples $S_1, S_2 \subseteq X \times \{0, 1\}$

$$f(S_1) \neq f(S_2) \text{ implies } \text{Dom}(f^*(S_1)) \neq \text{Dom}(f^*(S_2)).$$

Proof. Let $C$ be a finite concept class over a set $X$ and let $k \in \mathbb{Z}^+$. Assume that $(f, g)$ is an SCS of size $k$ over $C$. If $S \subseteq 2^{X \times \{0,1\}}$ is a sample labelled consistently with some concept in $C$ then $f(S) = \{x_1, x_2, \ldots, x_i\}$ where $0 \leq i \leq k$ and $x_1, x_2, \ldots, x_i \in S$. Next define a function $f^* : 2^{X \times \{0,1\}} \rightarrow (X \times \{0,1\})^*$ by

$$f^*(S) := (x_1, x_2, \ldots, x_i).$$

Let $R \in 2^{X \times \{0,1\}}$ and $L \in (X \times \{0,1\})^*$ be some ordering of the elements of $R$ then define

$$g^*(L) = g(R).$$

Then $\text{Dom}(f^*(S)) = f(S)$ and $|f^*(S)| \leq k$. Since $(f, g)$ is an SCS, the remaining properties required for $(f^*, g^*)$ to be an SBSCS of size $k$ are satisfied.

Now assume that $C$ has an SBSCS $(f^*, g^*)$ of size $k$ such that $\text{Dom}(f^*(S_1)) \neq \text{Dom}(f^*(S_2))$ for any two distinct samples $S_1, S_2 \subseteq X \times \{0,1\}$. Intuitively this just means that no compression sequence is a reordering of another compression sequence. So let $S \subseteq X \times \{0,1\}$ and note that for any compression set $R$, there exists a sequence
such that $R = \text{Dom}(L)$. Now define

$$f(S) := \text{Dom}(f^*(S)),$$

and

$$g(R) := g^*(L),$$

where $R = \text{Dom}(L)$ and $L \in (X \times \{0,1\})$.

Then it is easy to check that $(f,g)$ satisfies the properties of a sample compression scheme. \hfill \square

**Corollary 4.2.2.** Let $\mathcal{C} \subseteq 2^X$ be a concept class and let $k \in \mathbb{Z}^+$. If $\mathcal{C}$ has a sample compression scheme of size $k$ then $\mathcal{C}$ has a sequence-based sample compression scheme of size $k$.

**Theorem 4.2.3.** Let $\mathcal{C} \subseteq 2^X$ be a concept class over the finite set $X$ of VC dimension $d \in \mathbb{N}$. If $|X| = m \geq d$, then there does not exist a sequence based sample compression scheme of size $\frac{d}{\log ed + 1}$ for $\mathcal{C}$.

**Proof.** Let $S \subseteq 2^X$ be a set of unlabelled examples with $|S| = d$ that is shattered by $\mathcal{C}$. Assume that $f^* : 2^X \times \{0,1\} \rightarrow (X \times \{0,1\})^*$ is a sequence-based compression function for $S$. As $S$ is shattered by $\mathcal{C}$, all $2^d$ labellings of $S$ are possible sample sets for $\mathcal{C}$. Then there are at most

$$\sum_{i=0}^{k} \binom{d}{i} 2^i i! \leq \binom{d}{\leq k} 2^d d!$$
possible compression sequences of size \( k \) from \( S \).

So if we set \( |f^*(S)| = k \) then

\[
k \leq \frac{d}{\log ed + 1},
\]

if and only if

\[
d \geq k(\log ed + 1)
\]

\[
= k(\log ed + \log k - \log k + 1)
\]

\[
= k \log \frac{ed}{k} + k + k \log k
\]

\[
> \log \left( \frac{ed}{k} \right)^k + k + \sum_{i=1}^{k} \log i.
\]

Using Theorem 2.2.5 and the fact that

\[
\sum_{i=0}^{k} \binom{d}{i} 2^i i! \leq \binom{d}{\leq k} 2^k k!,
\]

we have that

\[
2^d > \left( \frac{ed}{k} \right)^k 2^k k!
\]

\[
\geq \binom{d}{\leq k} 2^d k!
\]

\[
\geq \sum_{i=0}^{k} \binom{d}{i} 2^i i!.
\]

57
Therefore if $S$ is shattered by $C$, then there does not exist a sequence-based reconstruction function $g^*$ for $S$ since there are less than $2^d$ compression sequences for $S$.

Theorem 4.2.3 has an added use by allowing us to derive bounds on the minimum size of an unlabelled sequence based sample compression scheme as well.

**Corollary 4.2.4.** Let $C \subseteq 2^X$ be a concept class of VC dimension $d \in \mathbb{N}$ over the finite set $X$. There does not exist a USBSCS of size $\frac{d}{\log ed+1}$ for $C$.

**Proof.** For any $C \subseteq 2^X$ such that there exists a USBSCS for $C$ we have that there must exist a SBSCS for $C$ with

$$|SBSCS(C)| \leq |USBSCS(C)|.$$  

Thus if there was a USBSCS of size $\frac{d}{\log ed+1}$ for $C$ then we would also have a SBSCS of size $\frac{d}{\log ed+1}$ for $C$ which would contradict Theorem 4.2.3.

**Theorem 4.2.5.** Let $C \subseteq 2^X$ be a maximum concept class over a finite set $X$ with VC dimension $d \in \mathbb{Z}^+$. If $|X| = m \geq d^2 2^{d-1}(d - 1)!$, then there does not exist a sequence-based sample compression scheme of size $d - 1$ for $C$.

**Proof.** Let $S \subseteq X$ be an unlabelled sample with $|S| = m \geq d^2 2^d!$. Then $|C|_S = \binom{m}{\leq d}$, meaning that $\binom{m}{\leq d}$ is the number of possible labellings of $S$. So for there to exist a sequence-based sample compression scheme $(f^*, g^*)$ for $S$ we must have $|\{f^*(S) \mid S \subseteq$
$|X| \geq \binom{m}{\leq d}$; otherwise $g^*$ would not be able to return all possible labellings of $S$, which would mean property (iii) of the SBSCS definition would not be satisfied for all samples $S' \subseteq X \times \{0, 1\}$.

Observe that for any sequence of $i$ elements there are $2^i$ labellings for this sequence and $i!$ orderings of the each labelled sequence. Therefore there are at most

$$\sum_{i=0}^{d-1} \binom{m}{i} 2^i i!$$

labelled compression sequences of size $d - 1$ for $S$.

To show that

$$\sum_{i=0}^{d-1} \binom{m}{i} 2^i i! < \sum_{i=0}^{d} \binom{m}{i} = \binom{m}{\leq d},$$

it suffices to show that

$$\sum_{i=0}^{d-1} \binom{m}{i} (2^i i! - 1) < \binom{m}{d}.$$

If $m \geq \left\lceil \frac{d}{2} \right\rceil$ then since

$$\sum_{i=0}^{d-1} \binom{m}{i} (2^i i! - 1) < d(2^{d-1}(d - 1)! - 1) \binom{m}{d - 1}$$

and

$$\binom{m}{d} = \binom{m}{d-1} \frac{m - d + 1}{d},$$

then we need

$$d(2^{d-1}(d - 1)! - 1) \binom{m}{d - 1} < \binom{m}{d - 1} \frac{m - d + 1}{d}.$$
This is true if and only if

\[ d^2 2^{d-1}(d-1)! - d^2 + d - 1 < m. \]

Therefore if we pick \( m \geq d^2 2^{d-1}(d-1)! \) we have that \( m \geq \lceil \frac{d}{2} \rceil \) and that

\[
\sum_{i=0}^{d-1} \binom{m}{i} 2^i i! < \binom{m}{d}.
\]

\[ \square \]

4.3 Compressing \( t \)-designs

**Definition 4.3.1.** A \( t-(m, k, \lambda) \) design, a \( t \)-design for short, is a pair \( (X, \mathcal{B}) \) where \( X \) is a set of \( m \) points and \( \mathcal{B} \) is a collection of subsets of \( X \) of size \( k \) called blocks with the property that every set of \( t \) points of \( X \) is in exactly \( \lambda \) blocks, we call \( \lambda \) the index of the design. If \( \lambda = 1 \) we call \( \mathcal{B} \) a Steiner system and denote this by \( \mathcal{B} := S(t, k, m) \).

**Example 4.3.2.** If \( |X| = 7 \) then the Fano plane is a \( 2-(7, 3, 1) \) design

\[
\mathcal{B} = \{ \{x_1, x_2, x_3\}, \{x_1, x_4, x_7\}, \{x_1, x_5, x_6\}, \{x_2, x_5, x_7\}, \{x_2, x_4, x_6\}, \{x_3, x_6, x_7\}, \\
\{x_3, x_4, x_5\} \} \subseteq 2^X.
\]

Visually, this is described by the diagram shown in Figure 4.1.
Here each point of $X$ labels a point on the plane and each $B \in \mathcal{B}$ corresponds to exactly one line in Figure 4.1.

Something to keep in mind is that by [23], for any $t$-$(m, k, \lambda)$ design $\mathcal{B}$ the quantity $\lambda$ is bounded by the integer $\binom{m-t}{k-t}$. In addition, if $\lambda = \binom{m-t}{k-t}$, then $\mathcal{B} = \binom{X}{k}$, where $\binom{X}{k}$ is called the complete design of order $k$.

**Theorem 4.3.3.** [14]/[23] Let $\mathcal{B}$ be a $t$-$(m, k, \lambda)$ design. If $0 \leq i + j \leq t$, then

$$\lambda_{i,j}(\mathcal{B}) = \frac{\lambda \binom{m-i-j}{k-i}}{\binom{m-t}{k-t}}$$

is an integer. Further, $\lambda_{i,j}$ counts the number of blocks in $\mathcal{B}$ that contain all of a set of $i$ points and none of a disjoint set of $j$ points.
Definition 4.3.4. Given a concept class \( C \subseteq 2^X \) we call

\[
\overline{C} = \{ X \setminus C : C \in C \}
\]

the **supplement** of \( C \).

Before using the supplement, we will state a few properties that will be useful in later results.

**Lemma 4.3.5.** [26] If \( C \) be a concept class over a set \( X \), then

\[ \text{VCD}(C) = \text{VCD}(\overline{C}) \]

**Lemma 4.3.6.** [23] If \( \mathcal{B} \) is a \( t-(m, k, \lambda) \) design, then \( \overline{\mathcal{B}} \) is a \( t-(m, m-k, \lambda_{0,t}(\mathcal{B})) \) design.

**Example 4.3.7.** Consider the Fano plane from Example 4.3.2, denoted \( \mathcal{B} \). If we consider the incidence matrix of \( \overline{\mathcal{B}} \) given by Table 4.1 and denoted \( I_{\overline{\mathcal{B}}} \).
Then we see that in each pair of points $x, z \in X$ there are exactly two concepts in $\overline{\mathcal{B}}$ that have both $x$ and $z$ labelled 1. Further if we consider $\{x_1, x_2, x_3\}$ we see that there is no concept in $\overline{\mathcal{B}}$ with each of these points labelled 1. This, together with the fact that each concept in $\overline{\mathcal{B}}$ has size 4 shows us that $\overline{\mathcal{B}}$ is a 2-(7, 4, 2) design as predicted by Lemma 4.3.6.

Additionally, we can compute the VC dimension of this concept class by restricting the incidence matrix $I_{\overline{\mathcal{B}}}$ to the set $\{x_1, x_2\}$ so that $I_{\overline{\mathcal{B}}}[\{x_1, x_2\}]$ is as given in Table 4.2.

It is then easy to see that $|\overline{\mathcal{B}}[\{x_1, x_2\}]| = 2^2 = 4$. Since $|\overline{\mathcal{B}}| = 7 < 2^3 = 8$ then we must have that $\text{VCD}(\overline{\mathcal{B}}) = 2$. Further using Lemma 4.3.5 and by noting that $\overline{\mathcal{B}} = \mathcal{B}$ it is clear that for the Fano plane $\mathcal{B}$ we have that $\text{VCD}(\mathcal{B}) = 2$. 

Table 4.1: The incidence matrix of the Fano plane

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X \setminus C_1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$X \setminus C_2$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$X \setminus C_3$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$X \setminus C_4$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$X \setminus C_5$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$X \setminus C_6$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$X \setminus C_7$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
\[
\begin{pmatrix}
C_1 & 0 & 0 \\
C_2 & C_3 & 0 & 1 \\
C_4 & C_5 & 1 & 0 \\
C_6 & C_7 & 1 & 1 \\
\end{pmatrix}
\]

Table 4.2: The restriction to \(x_1\) and \(x_2\) of the incidence matrix for the Fano plane

**Theorem 4.3.8.** If \(B\) is a \(t-(m, k, \lambda)\) design, then \(t \leq \text{VCD}(B) \leq k\).

**Proof.** If \(B\) is a \(t-(m, k, \lambda)\) design then the quantity \(\lambda_{i,j}(B)\) must be in \(\mathbb{N}\) for any \(0 \leq i + j \leq t\). In order to shatter a set of \(t\) points in \(B\) we need, for some \(t\)-set \(T \in X\), that, for any partition of \(T\) into two disjoint sets \(T = T_1 \cup T_2\) that \(B|_{T} = T_1\) and \(B|_{T} \cap T_2 = \emptyset\). Since \(\lambda_{i,j} \in \mathbb{N}\) it follows that we can find such a set if \(\lambda_{i,j}(B) \neq 0\) for all \(i + j = t\).

For ease of computation let us consider \(\overline{B}\). Since \(\text{VCD}(B) = \text{VCD}(\overline{B})\) we may do this without changing our result. By definition \(\overline{B}\) is a \(t-(m, m-k, \overline{\lambda})\) design where \(\overline{\lambda} = \lambda_{0,t}(B)\). Thus we will compute \(\lambda_{i,j}(\overline{B})\) for \(i + j = t\) and verify conditions where it is not equal to zero. So assume that \(\lambda_{i,j}(\overline{B}) = 0\) for some \(i + j = t\), then, since \(\lambda \geq 1\)
by assumption, the expression

\[ \lambda_{i,j}(B) = \lambda_{(m-t)}(m-k-t) = \lambda_{(m-k-t)}(m-t) \]

holds if and only if \( (m-t) = 0 \). However, this would imply that \( m - t < m - k \) but this holds if and only if \( t > k \), which is a contradiction.

\[ \square \]

**Theorem 4.3.9.** Every \( t-(m,k,\lambda) \) design is a \( (t-1)-(m,k,\lambda') \) design where

\[ \lambda' = \lambda \left( \frac{m - t + 1}{k - t + 1} \right) . \]

It is with this theorem in mind that we define for a \( t-(m,k,\lambda) \) design \( B \) the quantity \( t^{\text{max}} \) which is the largest value such that \( B \) is a \( t^{\text{max}}-(m,k,\lambda^*) \) design for some \( 1 \leq \lambda^* \leq \binom{m-t}{k-t} \). The definition of \( t^{\text{max}} \) along with the result from Theorem 4.1.9 allows for an easy corollary that will help to establish the significance of the compression scheme work that will follow.

**Corollary 4.3.10.** If \( B \) is a \( t^{\text{max}}-(m,k,\lambda) \) design then \( t^{\text{max}} \leq \text{VCD}(B) \leq k. \)

**Proof.** This follows from Theorem 4.3.8 since \( B \) is a \( t^{\text{max}}-(m,k,\lambda) \) design.

**Example 4.3.11.** If we once again consider the Fano plane defined in Table 4.1 and denote this \( B \) as before, then using Corollary 4.3.10 and using the geometric properties
of the design we have an alternate way to compute the VC dimension of $\mathcal{B}$. Notice first that for any pair of points in $X$, we can always find a line in the diagram that either, contains none of the points, contains exactly one or the other point or contains both points. This allows the observation that $\text{VCD}(\mathcal{B}) \geq 2$. However, since any set of 2 points defines a full line there are no 3 points such that there exists a set $B \in \mathcal{B}$ containing only two of the points while not containing the third, so $\text{VCD}(\mathcal{B}) = 2$. Also observe that $\mathcal{B}$ cannot be a $t$-design where $t \geq 3$ since by Corollary 4.3.10 we have $t^{\max} \leq \text{VCD}(\mathcal{B}) = 2$.

The following compression scheme uses the fact that there are $t!$ orderings of any $t$-set. If $\mathcal{B} \subseteq 2^X$ is a $t-(m,k;\lambda)$ design, with $\lambda \leq t!$, then for any $t$-set $T \subseteq X$ we can create a correspondence between the orderings of $T$ and the sets $B \in \mathcal{B}$ with $T \subseteq B$. The following USBSCS will make use of this fact.

Let $\mathcal{B}$ be a $t-(m,k,\lambda)$ design where $\lambda \leq t!$. Let $S \subseteq 2^{X \times \{0,1\}}$ be a sample labelled consistently with some $B \in \mathcal{B}$. If $T \subseteq X$ is a $t$-set let $(T)_1, (T)_2, \ldots, (T)_{t!}$ be all the possible orderings of $T$. Then since $X$ is an ordered set we may surjectively map each $(T_i)$ to some $B \in \mathcal{B}$ with $T \subseteq B$. For a $t$-set $D$, when an ordering of $D$ is mapped in this way to a set $B \in \mathcal{B}$ with $D \subseteq B$ we write $(D) = (D)_B$.

The $t$-design unlabelled sequence-based sample compression algorithm

The Compression Function $f^*$: To define the function $f^*$ we need to consider some cases.
Case 1: If $|1(S)| < t$ then $f^*(S) = (1(S))$ where $(1(S))$ is a sequence of points from $1(S)$.

Case 2: If $k \geq |1(S)| \geq t$ then choose some $D \subseteq 1(S)$ with $|D| = t$. We then let $f^*(S) = (D)_B$ where $B$ is a set in $B$ that is consistent with $S$.

The Reconstruction Function $g^*$: There are two kinds of input to $g^*$ that we need to distinguish. The input to this function is a sequence $(D) \in \binom{X}{t}^*$. 

i) When $|(D)| < t$ then $g^*((D)) = \{(x, y) \in X \times \{0, 1\} : y = 1$ if $x \in (D)$ and $y = 0$ if $x \notin (D)\}$

ii) When $|(D)| = t$, as established before, there exists some $B \in B$ such that $(D) = (D)_B$. We let $g^*((D)) = \{(x, y) \in 2^X \times \{0, 1\} : y = 1$ if $x \in B$ and $y = 0$ if $x \notin B\}$.

The following theorem shows that this choice of compression and reconstruction functions defines a unlabelled sequence-based sample compression scheme.

Theorem 4.3.12. If $B$ is a $t$-$(m, k, \lambda)$ design and if $\lambda \leq t^{\max}$, then $B$ has a USBSCS of size $t$.

Proof. Let $S_B$ be a sample consistent with a concept in $B \in B$ where $B$ is a $t$-$(m, k, \lambda)$ design where $\lambda \leq t$! and let $(f, g)$ be a USCS of size $d$ for $B$. To show that our compression function is a USBSCS of size $t$ for $B$ we proceed by verifying that each of the properties $(i)$ to $(iv)$ in the definition of a USBSCS are satisfied.
i) For each case in the definition of the compression function $f^*$ in this compression scheme we have $\text{Dom}(f^*(S_B)) \subseteq X(S_B)$.

ii) Further because the compression function only compresses to sets of size at most $t_{\text{max}}$ we have that $|f(S_B)| \leq t_{\text{max}}$.

iii) To verify that $S_B \subseteq g^*(f^*(S_B))$ we first look at the case when $|1(S_B)| < t$.

   In this case $g^*(f^*(S_B)) = g^*((1(S_B))) = S_1(S_B)$ where $|S_1(S_B)| = |X|$ so that $S_B \subseteq S_1(S_B)$. The other case is where $k \geq |1(S_B)| \geq t$, in this case we have that the $t$-sequence $f^*(S_B)$ has an ordering that corresponds to a $B' \in \mathcal{B}$ with $f^*(S_B) \subseteq B'$ and $1(S_B) \subseteq B'$ so that $g^*(f^*(S_B)) = \{(x, y) \in 2^X \times \{0, 1\} : y = 1 \text{ if } x \in B' \text{ and } y = 0 \text{ if } x \notin B'\}$. Thus $S_B \subseteq g^*(f^*(S_B))$.

iv) Finally by definition we see that no $x \in X$ has that both $(x, 1), (x, 0) \in g^*(f^*(S_C))$ for any sample $S_C$ so that property (iv) is satisfied.

Example 4.3.13. As an application of the Design Compression Scheme, let $\mathcal{B}$ be the Fano plane defined in Example 4.3.2, then each 2-subset of $X$ is contained in exactly one block in $\mathcal{B}$ we can define an unlabelled compression scheme of size 2 for $\mathcal{B}$ by compressing $1(S)$ to some 2-subset $D \subseteq 1(S)$ for any samples $S$ with $|1(S)| \geq 2$. The reconstruction function simply returns the set $\{(x, y) : y = 1 \text{ if } x \in B \in \mathcal{B} \text{ and } y =$
0 elsewhere} where $B \in \mathcal{B}$ with $D \subseteq B$. If $|1(S)| < 2$ we just compress to the set $1(S)$ and return the set $\{(x, y) : y = 1 \text{ if } x \in 1(S) \text{ and } y = 0 \text{ elsewhere}\}$.

Notice that here we do not use any other orderings of the 2-sequences to compress the Fano plane. Since no orderings of sequences were used, we can treat this USBSCS as a USCS. It turns out that this property can be generalized to all Steiner systems.

**Corollary 4.3.14.** If $\mathcal{B} = S(t, k, m) \subseteq 2^X$ is a Steiner system then $\mathcal{B}$ has a USCS of size $t$.

**Proof.** Let $\mathcal{B} = S(t_{\max}, k, m) \subseteq 2^X$ be a Steiner system. This implies that $\lambda = 1 \leq t!$. Using Theorem 4.3.12 we can conclude that $\mathcal{B}$ as a USBSCS of size $t$ using the Design Compression Scheme. Since for this design $\lambda = 1$, if $D$ is a compression sequence under this scheme and if $D'$ is another compression sequence, with $|D'| = |D|$, then Dom($D$) $\neq$ Dom($D'$). So to transform this USBSCS into a USCS we only need to replace each compression sequence $D$ with its corresponding domain set Dom($D$).

By the proof of Theorem 4.3.12, this new compression scheme satisfies the properties required to be a USCS since, for our purposes, we can treat compression sets under our USCS as sequences under a specified ordering in our USBSCS. \hfill \Box

**Corollary 4.3.15.** If $\mathcal{B} = S(t_{\max}, k, m)$ is a Steiner system then $\text{VCD}(\mathcal{B}) = t_{\max}$.

**Proof.** From Corollary 4.3.10 we can establish that $\text{VCD}(\mathcal{B}) \geq t_{\max}$. Further, since $\mathcal{B}$ is a $t_{\max}$-($m, k, 1$) design by Corollary 4.3.14, $\mathcal{B}$ must have a USCS of size $t_{\max}$. But this implies that $\text{VCD}(\mathcal{B}) \leq t_{\max}$ so that $\text{VCD}(\mathcal{B}) = t_{\max}$. \hfill \Box
Chapter 5

Conclusion

5.1 Summary

To conclude, this thesis has developed a number of results on labelled and unlabelled sequence-based sample compression. These results include an unlabelled sequence-based sample compression scheme of size \( \lfloor \frac{n^2}{7} \rfloor \) as well as showing that for small instance space size, it is possible to create an unlabelled sequence-based sample compression scheme for maximum classes that has size \( d - 1 \). Also some bounds on the minimum sizes of labelled sequence-based sample compression schemes were attained. In addition we have introduced a new way to define sample compression schemes, one that extends other generalizations of sample compression schemes. Finally we have explored the connection between unlabelled sample compression schemes, both sequence-based and set-based, and combinatorial design theory. This includes demonstrating that it is possible to find an unlabelled sequence-based sample compression
scheme of size $t$ for a $t$-$(m, k, \lambda)$ design, provided $\lambda \leq t$.

5.2 Limitations and Open Questions

In the process of investigating sequence-based sample compression schemes as well as their relation to design theory, there are a number of questions that arise. Here we will discuss some of these questions. An obvious one to ask, is related to the extension of Conjecture 2.5.6 to sequences, namely, for any concept class $\mathcal{C}$ does there exist a sequence-based sample compression scheme of size $O(d)$ where $\text{VCD}(\mathcal{C}) = d$?

Although we have shown that the existence of an unlabelled sequence-based sample compression scheme for a concept class $\mathcal{C}$ implies the PAC-learnability of $\mathcal{C}$, the method we used to show this was just by simple observation of the fact that USBSCS’s are just special cases of $\{n_i\}_{i=0}^k$-copy sample compression schemes. This left us with a sample complexity for the PAC learning of $\mathcal{C}$ that is somewhat higher than if we wanted to PAC learn $\mathcal{C}$ with the standard sample compression schemes. An interesting approach would be to consider if there are further improvements on the sample complexity of PAC learning a concept class with a sequence-based sample compression scheme, of either the labelled or the unlabelled variety.

In the design theory section of this thesis we establish an unlabelled sequence-based sample compression scheme of size $t$ for $t$-$(m, k, \lambda)$ designs. However, this is a somewhat restrictive compression scheme as we require that $\lambda \leq t!$. An open question
along these lines is whether it is possible to remove this restriction on $\lambda$. Further, our work has shown that $t^{\text{max}} \leq \text{VCD}(\mathcal{B})$ for any $t^{\text{max}}$-design, and for Steiner systems, this bound is tight, but it is not known whether $t^{\text{max}} = \text{VCD}(\mathcal{B})$ in general.

Finally, although we obtained many results on sequence-based sample compression, these results used the fact that there were more objects available to compress to when improving bounds over those already known using set-based sample. An interesting avenue of research might be to see if there were some structural properties of sequences that can be exploited to improve compression bounds.
Bibliography


